The effects of local, catchment and climatic factors on the reliability of microbial bioindicators: diatoms in fluvial ecosystems

VIRPI PAJUNEN

ACADEMIC DISSERTATION
To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public examination in Auditorium XII, University main building, on 2 March 2018, at 12 o'clock noon.

DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY A59 / HELSINKI 2018

Cover photo: Virpi Pajunen

Author´s address: Virpi Pajunen
Department of Geosciences and Geography
P.O.Box 64
00014 University of Helsinki
Finland
virpi.pajunen@helsinki.fi

Supervised by: Professor Janne Soininen
Department of Geosciences and Geography
University of Helsinki, Finland

Professor Miska Luoto
Department of Geosciences and Geography
University of Helsinki, Finland

Reviewed by: Professor Elie Verleyen
Department of Biology
Ghent University, Belgium

Professor Donald F. Charles
Department of Biodiversity, Earth & Environmental Science
Drexel University, U.S.A

Discussed with: Professor John P. Smol
Department of Biology
Queen's University, Canada

ISSN-L 1798-7911
ISSN 1798-7911 (print)
ISBN 978-951-51-2943-7 (paperback)
ISBN 978-951-51-2944-4 (pdf)
http://ethesis.helsinki.fi

Painosalama Oy
Turku 2018

Pajunen V., 2018. The effects of local, catchment and climatic factors on the reliability of microbial bioindicators: diatoms in fluvial ecosystems. *Department of Geosciences and Geography* A59. 40 pages and 5 figures.

Abstract

The ongoing climate change and increasing anthropogenic pressure threaten the biodiversity on Earth. Elevated temperatures, changes in precipitation and intensive land use alter ecosystems and such changes are prone to escalate in the northern regions, especially in freshwater ecosystems. Species must thus respond to these changes by adaptation or adjusting their distributional ranges. Information about the effects of climate on the distributional patterns of diverse aquatic microorganisms has yet largely been lacking. This is a drawback as microbial species in freshwaters play crucial roles in ecosystem functioning as well as in environmental monitoring. Thus, it is necessary to disentangle the main drivers of microbial species distributions in order to predict the responses of freshwater communities to future environmental change and to ensure the accurate determination of the ecological status of ecosystems.

This doctoral thesis aims to investigate the relative roles of climate, catchment properties and local environmental factors in the occurrence of the important freshwater micro-organisms both at species and community levels. This study, conducted at a regional scale (c. 1000 km), concentrates on unicellular stream diatoms, which are widely used in biomonitoring. In detail, the study seeks to reveal (1) whether diatom species distributions are influenced by climatic factors or solely driven by local environmental variables, (2) whether the importance of the factors governing species distributions varies along the anthropogenic land use gradient, (3) the pathways and the effects of climate, land use and the most important local environmental variables on diatom diversity and community composition, and (4) the ability of diatom assemblages to predict climatic and local environmental variables.

The results showed that climatic factors are important drivers of stream diatom distributions and their influence may even outcompete the effects of local environmental variables. However, the relative importance of the factors governing diatom distributions varied along the anthropogenic land use gradient and among species. Climate was the main driver of species distributions in pristine environments, whereas local environment was more important in human impacted streams. Climatic and catchment scale factors affected stream diatoms mainly via indirect pathways, for example, through catchment productivity and nutrient availability. Species richness was mainly influenced by energy and nutrient availability. Conductivity, which was strongly related to anthropogenic land use, was a key factor influencing community composition and uniqueness, but also species distributions especially in human impacted streams. Unique communities with high conservation value and low species richness were detected in harsh, low-nutrient conditions in northern Finland. Diatom assemblages were also found to be reliable predictors of both climatic and local environmental factors indicating their robustness as environmental proxies and bioindicators. Highly suitable indicator species were identified for water chemistry variables but also for certain climatic conditions.

This thesis contributes to the spatial research of aquatic micro-organisms as it brings a novel evidence of the biogeographical patterns of microbial species. This study revealed that climate, one of the fundamental drivers of species distributions on Earth, governs also the occurrences and abundances of stream diatoms even at regional scales. However, it is important to acknowledge that the effects of the most essential climatic and environmental factors influencing diatom species may be context dependent and vary along the anthropogenic land use gradient. The ongoing climatic and subsequent environmental change may further complicate the species responses towards environmental factors. From an applied perspective, this study confirmed the reliability of stream diatom assemblages as bioindicators. However, diatom responses towards novel environmental conditions need to be reevaluated to assure their accuracy also in the future.

Acknowledgements

”*I have no special talents. I am only passionately curious*”, Albert Einstein.

Perhaps, my never-ending curiosity and perseverance have led me to this moment of a great achievement. It wasn't always easy and many times I found myself in times of trouble and despair. But luckily, I got by with a little help from my friends, co-workers, and of course, my supervisors.

I am deeply grateful for having the best supervisors one could hope for: professors Janne Soininen and Miska Luoto. They both have advised and supported me from the very beginning when I was still pursuing my dream and desperately seeking for funding to start my PhD thesis. I have been privileged to work with such efficient and enthusiastic experts in their field. I can truly say that I have learned from the best.

I wish to thank the preliminary examiners, professors Elie Verleyen (Ghent University) and Don F. Charles (Drexel University), for the insightful and supporting comments on the manuscript of this PhD thesis.

I could not have hoped for a better working environment and co-workers. I am greatly indebted to Sandra Meyer, Anette Teittinen and especially to Jenny Jyrkänkallio-Mikkola for accompanying me in the adventurous field work. Together we conquered the most beautiful stream sites and the most dreadful ditches, admired amazing landscapes and marvelled remote (and sometimes creepy) settlements, survived the occasional attacks by bloodsucking creatures and vicious riparian plants, and shared all the ups and downs of this work. Thank you, I had a real blast.

At the Department of Geosciences and Geography, I have been privileged to work close to many talented scientists. Special thanks to Tua Nylén, Maija Taka, Juha Aalto, Konsta Happonen, Julia Kemppinen, Mikko Korpela and professor Mathieu Cusson. We did not share only a small working space but also a great sense of humour and many common interests, which made even the dullest and darkest days more endurable. I also received much precious advice and help from these guys, of which I am deeply grateful. I owe a very special thanks to Maija who had always time to help me and had a remarkable gift for finding an answer to any question I thought of asking. I extend my gratitude to Juhani Virkanen, who shared his expertise concerning the questions related to field and lab work, and Arttu Paarlahti without whom I would have smashed my computer to the wall several times during my PhD work.

On a personal note, I want to thank my family and friends. I am deeply grateful for my parents, Tuula and Raimo, who have always supported me and encouraged me to pursue all my ambitions. I thank my sister Erja and brothers, Harri and Sami, from whom I have learned practicality, great imagination and determination by following their examples. I am deeply grateful for my best friends Nina and Suvi, who I know will always stand by me thought I have occasionally been buried at work. I wish to acknowledge my mother-in-law Raija who kindly helped me to improve the grammar in this PhD thesis.

Finally, I want to thank my husband Lassi, for his love, support and patience, and my lovely daughters Pinja and Nella, who seem to have grown quite normal despite all this.

For funding I wish to thank Maj and Tor Nessling Foundation, Nordenskiöld-Samfundet and Emil Aaltonen Foundation.

This PhD thesis is dedicated to my dearest friend Satu Lampinen (1982 – 2014). She was so excited and proud when I started my PhD thesis. Unfortunately, she cannot be here to see my thesis finished as she suddenly passed away in 2014. I am forever grateful for her unconditional support and friendship.

In Helsinki, January 24th 2018

Virpi Pajunen

Contents

List of original publications

This thesis is based on the following publications:

I **Pajunen, V.**, Luoto, M., Soininen, J. 2016. Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography* 25, 198-206.

II **Pajunen, V.**, Jyrkänkallio-Mikkola, J., Luoto, M., Soininen, J. 2018. Are drivers of microbial bioindicators context dependent in human impacted and pristine environments? Submitted manuscript.

III **Pajunen, V.**, Luoto, M., Soininen, J. 2017. Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities. J*ournal of Biogeography* 44, 2376-2385.

IV **Pajunen, V.**, Luoto, M., Soininen, J. 2016. Stream diatom assemblages as predictors of climate. *Freshwater Biology* 61, 876-886.

The publications are referred to in the text by their roman numerals.

Authors' contributions

I The study was planned jointly by V. Pajunen, M. Luoto and J. Soininen. The data were collected and analysed by J. Soininen. V. Pajunen conducted the data preparations and the statistical analyses except the SDMs. The SDMs were conducted by M. Luoto and V. Pajunen. V. Pajunen prepared the manuscript and it was commented by J. Soininen and M. Luoto.

II The study was planned jointly by V. Pajunen, M. Luoto and J. Soininen. The data were collected and analysed by J. Soininen, V. Pajunen and J. Jyrkänkallio-Mikkola. V. Pajunen conducted the data preparations, catchment analyses and the statistical analyses except the SDMs. The SDMs were conducted by M. Luoto. V. Pajunen prepared the manuscript and it was commented by J. Soininen, M. Luoto and J. Jyrkänkallio-Mikkola.

III The study was planned jointly by V. Pajunen, M. Luoto and J. Soininen. The data were collected and analysed by J. Soininen. V. Pajunen conducted the data preparations, catchment analyses and the statistical analyses except the SEMs. The SEMs were conducted by M. Luoto and V. Pajunen. V. Pajunen prepared the manuscript and it was commented by J. Soininen and M. Luoto.

IV The study was planned jointly by V. Pajunen, M. Luoto and J. Soininen. The data were collected and analysed by J. Soininen. V. Pajunen conducted the data preparations and the statistical analyses except the inference modelling. The inference models were conducted by M. Luoto. V. Pajunen prepared the manuscript and it was commented by J. Soininen and M. Luoto.

Abbreviations

Anthro	anthropogenic land use
AUC	area under the receiver operating characteristics curve
BRT	boosted regression tree; aka generalized boosted model (GBM)
GAM	generalized additive model
GDD	growing degree days
GLM	generalized linear model
LCBD	local contribution to beta diversity
MAT	modern-analogue technique
NMDS	non-metric multidimensional scaling
PCA	principal component analysis
PRECS	precipitation sum from May to September
RDA and pRDA	redundancy analysis and partial redundancy analysis
r^2	coefficient of determination
RF	random forest
RMSEP	root-mean-square error of prediction
SDM	species distribution model
SEM	structural equation model
TP	total phosphorus
TSS	true skill statistics
WA	weighted averaging
WAB	water balance
WA-PLS	weighted averaging partial least squares

List of figures

List of tables

1 Introduction

Climate is a fundamental factor governing species distributions on Earth (Davis and Shaw, 2001; Hughes, 2000; McCarty, 2001; Walther *et al*., 2002), but during the last centuries, human actions have become increasingly influential in altering environmental conditions and processes. Due to the ongoing climate change, the rising mean temperatures and the changes in precipitation are modifying the environmental conditions for current biota (IPCC, 2014). Furthermore, ecosystems are affected by other anthropogenic stressors, such as changes in land use, which may lead to an environmental degradation and homogenization of biota (Rahel, 2002; Donohue *et al*., 2009, Dar and Reshi, 2014). The changes are escalating in boreal and arctic regions and freshwater ecosystems are especially susceptible to these changes (Heino *et al*., 2009). Species' responses to the projected environmental change depend on their individual traits and the factors driving their distribution (Parmesan and Yohe, 2003). Species with restricted ranges are particularly vulnerable to changes in their habitat, because they may not be able to adapt to the new habitat conditions (Parmesan, 2006).

1.1 Stream habitat: the hierarchy of multiscale environmental factors

Among the freshwater ecosystems, fluvial waters, i.e. rivers and streams, are an essential part of global hydrological cycle as they transport water from the land to the sea together with soils, nutrients and other materials (Allan and Castillo, 2007). They also provide important ecosystem services for humans, such as clean water supply and resources for industry, and the demand of these services increases together with the population growth (Baron *et al.*, 2002). Ecologically diverse and functionally intact fluvial systems are more likely to buffer the ongoing and projected environmental change (Chapin *et al*., 1997; Baron *et al*., 2002).

Fluvial waters are characterized as complex and highly connected dendritic systems with unidirectional flow, and high frequency and intensity of environmental fluctuations (Allan and Castillo, 2007). The flow of water, energy and substances in streams vary not only in time but also between individual streams due to the interplay of numerous factors: the amount and composition of precipitation, the paths of water flowing through the catchment, substances derived from bedrock, soils and terrestrial vegetation, and the effect of human alterations (Moss, 1998). The continuous variability in stream physical conditions from headwaters to downstream produces a corresponding continuum of biological responses to the different habitats available (Vannote *et al.*, 1980).

The hierarchical approach to the stream habitat classification (Frissell *et al*., 1986) is based on an assumption that stream communities are governed jointly by the characteristics of the stream habitat and the pool of species available for colonization, and further, the stream habitat is determined by the stream catchment. Ultimately, the development and physical features of a stream system are dictated by geology, history and climate (Frissell *et al*., 1986; Biggs, 1996; Stevenson, 1997; Snelder and Biggs, 2002). Thus, streams are hierarchically and spatially nested systems, where the larger scale systems constrain the smaller scale systems within.

The species occurrences at a locality can result from environmental filtering, which operates at multiple spatial and temporal scales: namely large scale (comprising historic, climatic and evolutionary factors), dispersal (comprising regional species pool richness and dispersal distance) and environmental filter (comprising

habitat features) (Zobel, 1997; Hillebrand and Blenckner, 2002). Each filter limits species colonization from the available species pool (for instance, global or regional pool), and thus, only the species that have overcome these constraints are able to live and reproduce in the local habitat. In streams, a modification of this principle can be applied to describe the spatiotemporal filters operating in the river continuum (Fig. 1). These filters consist of spatial filters that operate at a large scale (consisting of climate and geology factors) and a catchment scale (consisting of land use and land cover factors), and finally local scale filters consisting of biotic and abiotic factors in the local habitat, which can be reach, riffle, pool or microhabitat, depending on the size of the organisms (Frissell *et al.*, 1986).

1.1.1 Large scale factors

The main characteristics of stream hydrology, channel shape and water chemistry result from climate, geology, topography and the catchment properties including human actions (Allan and Castillo, 2007). The ultimate determinants, climate and geology, interact with each other (i.e. weathering and erosion) and together they influence the factors operating at a catchment scale, such as land cover, as well as at the local scale (for example, temperature and substratum) (Stevenson, 1997; Fausch *et al*., 2002). The key aspects of climate are temperature (i.e. solar energy) and water (i.e. the hydrological cycle). Temperature is a fundamental requirement for life as it influences metabolism and thus vital ecological functions such as photosynthesis, respiration and growth (Brown *et al*., 2004). In addition to precipitation, life on Earth is distributed mainly according to the temperature demands of organisms (Cox *et al*., 2010).

Precipitation is the ultimate input of water to the stream system, and additionally, the subsequent run-off transports materials and substances from the catchment to the stream (Moss, 1998; Allan and Castillo, 2007). Streams can be classified based on the frequency and pathway (via run-off, groundwater or both) of the water input as intermitted (experiencing seasonal droughts) or perennial (year-around base flow) streams (Allan and Castillo, 2007). Stream flow velocity is a consequence of topography and precipitation. Individual streams have their own natural flow regime based on the geo-climatic features of their catchments (Poff *et al*., 1997). The relative amount of surficial run-off and water restrained in the catchment or filtrated into the aquifer depends on the vegetation, soil type and land use, for instance the amount of impervious surfaces (Allan and Castillo, 2007). Thus, the type of land cover together with topography and storm events influence the frequency and magnitude of flow related disturbances.

Geology and geomorphology affect the catchment features, topography, and both physical and chemical properties of the stream (Moss, 1998; Allan and Castillo, 2007). The patterns in large scale geomorphology can function as dispersal barriers to species, for example, mountains and oceans (Cox *et al*., 2010). These geological formations can also affect local climatic conditions, shown as altitudinal changes in the temperature, for instance. Additionally, history (for example, plate tectonics and past climatic changes) has a substantial role in the present-day biogeographical patterns (Cox *et al*., 2010). The effects of the evolutionary history in each geographical region are reflected in the biome and landscape, which influence the properties of the stream systems.

Furthermore, human actions can operate at the large scale level, as airborne pollutants from industry and traffic (such as atmospheric N and S depositions) can spread widely from their origins and the deposition is further enhanced by climate

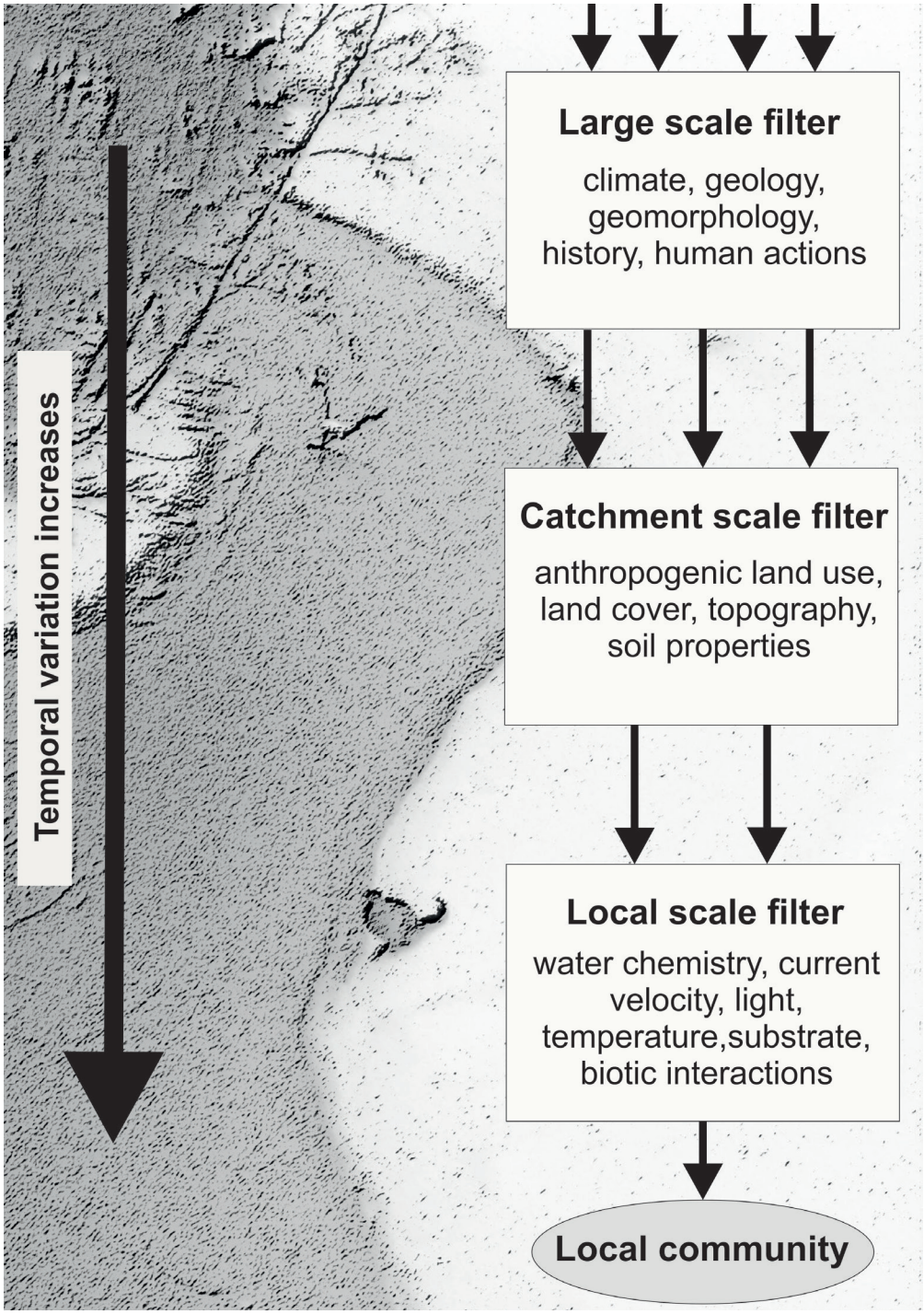


Figure 1 A schematic presentation of the spatiotemporal filters affecting stream diatom occurrences at a stream locality. The number of arrows pointing at each filter represent the size of the species pool available for colonization. The temporal variation in the stream conditions increases as the spatial scale decreases. Adapted and modified after Frissell et al. (1986) and Hillebrand and Blenckner (2002).

change (Kryza *et al*., 2012). Especially the acidification of freshwaters has become a problem in for example northern Europe where the streams have a naturally poor buffering capacity (Planas, 1996). The difficulty in assessing the effects of large scale factors on stream habitat and biota is a consequence of the complex pathways they operate through, i.e. indirect effects through intermediate variables (Stevenson, 1997).

1.1.2 Catchment scale factors

Local conditions in the streams depend on the upstream influence due to the continuous flux of materials and substances in the water flowing from upstream and the catchment (Moss, 1998; Allan and Castillo, 2007). The catchment serves as an important source of energy as many streams receive most of the carbon and nutrients as allochthonous inputs. However, the importance of terrestrial energy input varies among land cover: in shaded forested catchments most of the stream carbon originates from the catchment, but in unshaded regions, the major energy source can be in-stream primary production. In pristine streams, the base flow chemical concentration reflects catchment geology and land cover (Allan, 2004; Rothwell *et al.*, 2010; Varanka and Luoto, 2012). For instance, wetlands and coniferous forests are typical sources of humic substances and thus naturally acidic (Eshleman and Hemond, 1985) and such conditions can lead to stream brownification (Evans *et al.*, 2005). The strength of the catchment's influence may correspond to the size of the catchment area and thus become stronger downstream in the river continuum (Tudesque *et al*., 2014; Levesque *et al.*, 2017). However, the complexity of indirect catchment effects also increases downstream.

As the valley rules the stream (Hynes, 1975) and the human activities rule the valley (Allan, 2004), land use has a strong influence on stream habitats. Changes in the catchment land use will reflect in water stream chemistry, the flow regime and biota (Allan, 2004; Foley *et al.*, 2005; Varanka and Luoto, 2012). Pollutants, excessive nutrients and sediments derived from anthropogenic sources (such as agriculture, forestry, peat mining and urbanization) have a negative effect on stream water quality (Kolpin *et al*., 2002; Sutherland *et al*., 2002; Taka *et al*., 2017). Human actions can also affect the physical factors in streams, for example through vegetation removal, impervious surfaces and drainage systems, which impact the flow regimes and can result in flooding (Changnon and Demissie, 1995).

1.1.3 Local scale factors

The stream reach can be a very heterogeneous habitat due to woody debris, variation in rock grain size, erosion and possible human alterations (Palmer and Poff, 1997; Allan, 2004). Shading may differ in small areas owing to the variation in the amount of riparian vegetation, macrophytes or debris (Allan and Castillo, 2007). As a consequence of variation in stream morphology along the reach, current velocity differs between smaller sections of the channel forming distinctive habitats: pools and riffles (Frissell *et al*., 1986; Allan and Castillo, 2007). Pools can be seen as depositional habitats, where water movement is minimal, and riffles as erosional zones with a fast current velocity.

At even a smaller scale, microhabitats occur on a specific substrate (for instance, on rocks, sediment or plants) having relatively homogeneous water depth and current velocity due to their small size (Burkholder, 1996; Stoodley *et al.*, 2002; Battin *et al.*, 2007). These microhabitats inhabit the growth forms of resistant species such as algae, bacteria, fungi, bryophytes and meiofauna (Burkholder, 1996; Besemer, 2015). The chemical conditions of these habitats can differ from those of the overlaying water be-

cause of a boundary layer protecting the microbial community (Burkholder, 1996; Stoodley *et al.*, 2002; Battin *et al*., 2007). At the local scale, stream biota is influenced by various physical (i.e. temperature, light conditions and current), chemical (i.e. water and substrate chemistry) and biological factors (competition and grazing) in their immediate surroundings (Stevenson, 1997).

1.1.4 Temporal factors

Streams can be extremely disturbed ecosystems and highly variable in time mainly due to the variation in the flow (Allan and Castillo, 2007). Large disturbances, such as ice age, which have formed the landscape, have a long-lasting effect on the present stream conditions and species distributions (Cox *et al*., 2010; Vyverman *et al*. 2007; Vilmi *et al*., 2017). Likewise, even the past land use can have an imprint on a stream system for decades (Maloney and Weller, 2011). Streams in boreal and arctic regions are subjected to a large seasonal variation in the light regime and climatic conditions. Spring floods can constitute to a large portion of yearly nutrient loads (Buck *et al*., 2004). Storm events and drought, but also human actions, cause divergent stream conditions, and the frequency of these disturbances in a system affects the composition of stream assemblages (Lake, 2000; Schneck *et al*., 2017). The timing and magnitude of the last disturbance determine the successional stage of the biota (Biggs, 1996; Smucker and Vis, 2013).

1.2 The ecology and biogeography of benthic stream diatoms

Diatoms (Bacillariophyceae) are microscopic unicellular algae living in a wide variety of moist environments (Round *et al*., 1990). In streams, diatoms mainly live in benthos as members of biofilm, either on the sediment or attached on the surfaces of rocks or plants (Burkholder, 1996; Besemer, 2015). Diatoms are important primary producers in the stream food webs. They compete for resources, such as light, space and nutrients, with each other, other benthic algae, such as green algae, bacteria and mosses (McCormick, 1996). Diatoms are an important and/or preferable food source for stream herbivores, such as macroinvertebrates (for example snails and small insects), due to their nutritional value (e.g., high-energy lipids) (Smol and Stoermer, 2010). Benthic diatoms are mainly photoautotrophic, i.e. they photosynthesize, yet also facultative heterotrophy, i.e. additional ability to synthesize organic compounds, occurs among diatom species (e.g., Lewin and Lewin, 1960; Oliveira and Huynh, 1990). Diatoms are widely studied, and hence, a variety of factors have been found to influence their distribution and abundance (Fig. 2).

1.2.1 Species richness

Diatoms are a very species-rich group, with an estimated 24,000 – 200,000 species (Smol and Stoermer, 2010). The species identification has traditionally been based on the unique morphologies of the silica cell wall, but the development of molecular techniques will presumably increase the knowledge of diatom diversity in the near future (Zimmermann *et al*., 2015; Malviya *et al.*, 2016; Rimet *et al*., 2016). Local diatom species richness is strongly affected by the size of the regional species pool (Passy, 2009) and the amount of available habitat space corresponding with the environmental heterogeneity in the stream system. For now, the knowledge of the factors and processes controlling the diatom diversity in streams is not all-inclusive.

The latitudinal diversity gradient presents a universal decline in species richness towards the poles with only a few exceptions (Hillebrand, 2004). Studies of stream diatoms indicate that there is not a uniform response of diatom spe-

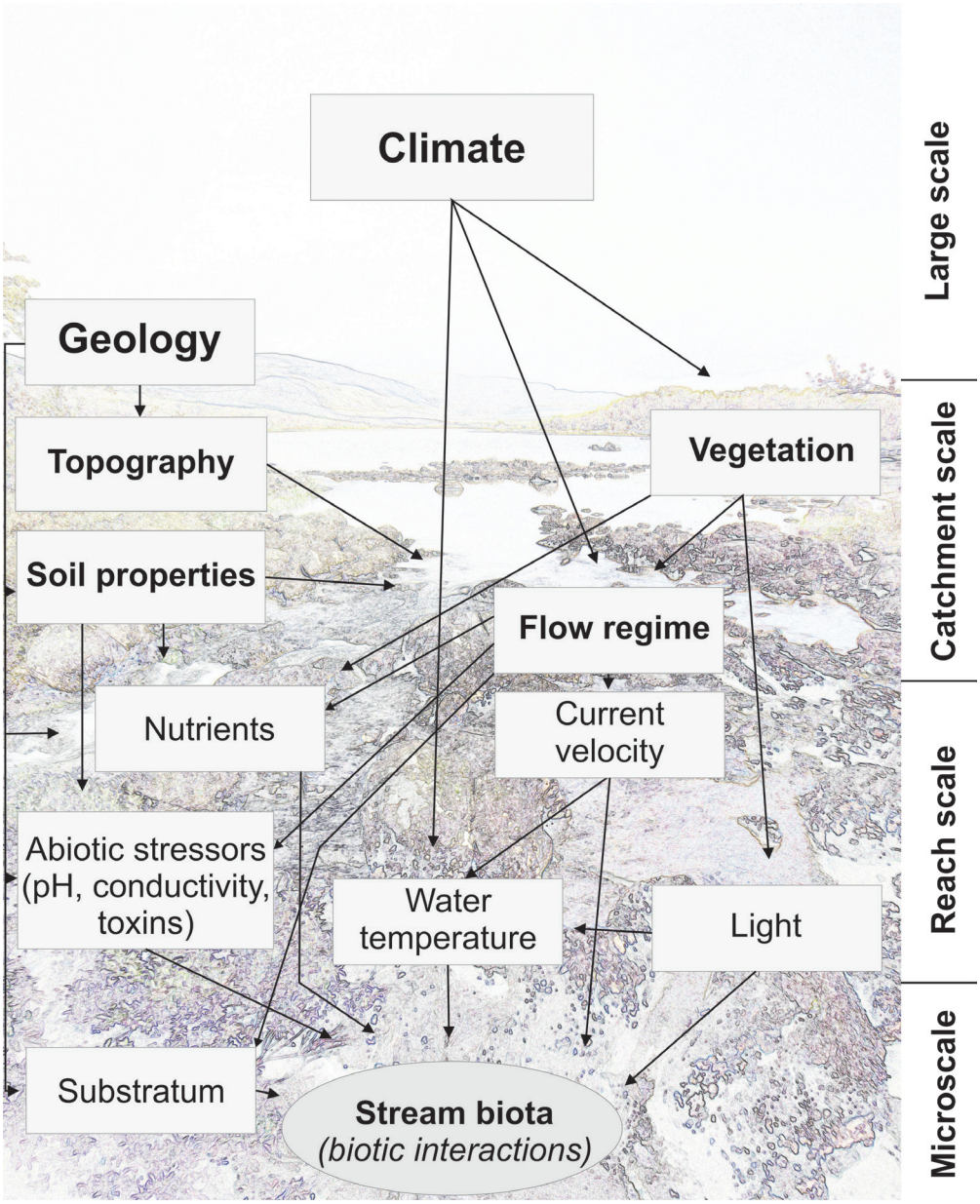


Figure 2 A simplified diagram demonstrating the hierarchical structure and interrelations of the main factors affecting stream biota (e.g., benthic diatoms) at multiple spatial scales. The local physiochemical factors in a stream operate at both reach and microscale. Biotic interactions occur in stream communities. Font sizes and thickness are scaled to match the spatial scale of the influence. Adapted and modified after Stevenson (1997).

cies richness to latitude and the responses vary across the study scale (Passy, 2010; Soininen *et al.*, 2016). Soininen *et al.* (2016) found a slight latitudinal increase in richness globally, yet Passy (2010) found a unimodal response at a continental scale. The possible effect of climate on species richness is relevant as climate changes dramatically from the equator to the poles

and with increasing altitude. In a regional study, Jyrkänkallio-Mikkola *et al*. (2017) found a positive response of stream diatom richness to growing degree days, a measure of energy availability. Correspondingly, Wang *et al*. (2017) observed either unimodal or decreasing pattern of diatom richness along elevational gradients. However, the latitudinal, altitudinal and climatic patterns of stream diatom richness at smaller spatial scales may not be linked to the temperature but rather to the corresponding patterns of catchment and stream properties, such as land use and water chemistry.

Diatom species richness seems to be affected by the local habitat conditions (for example, resource and niche availability), the size of the regional species pool and regional catchment characteristics (Passy, 2009; 2010). Diatom species richness has been positively related to the amount of wetlands (Passy, 2010; Pound *et al.*, 2013) or agricultural land use in the catchment (Jyrkänkallio-Mikkola *et al*., 2017). This indicates that catchment land use can be an important source of macro- and micronutrients (such as N, P, Mn and Fe) associated with increased niche-dimensionality and thus species richness (Liess *et al*., 2008; Passy, 2008; 2009; 2010; Johnson and Angeler, 2014). In fact, moderate nutrient enrichment can increase species richness (Lobo *et al*., 1995; Jüttner *et al*., 1996), but intensive land use can decrease species richness due to stream degradation by nutrient enrichment and toxicants (Yu and Lin, 2009; Teittinen *et al.*, 2015). Similarly, Passy (2009; 2010) found that the amount of forest cover negatively correlated with species richness presumably as a result of light depression and nutrient retention. This reasoning is supported by the fact that diatom species richness may increase with light intensities (Liess *et al*., 2008).

Species richness may decrease due to interspecific competition when dominant competitors exclude inferior competitors from the habitat. The Intermediate Disturbance Hypothesis (Connell, 1978) postulates that species richness peaks at intermediate levels of disturbance because of trade-offs between resource competition and re-colonization after a disturbance. A unimodal response to the variation in current velocities (physical disturbance) and grazing (biological disturbance) is a documented pattern in stream benthic algae richness (Stevenson *et al.*, 1996, and the references therein).

1.2.2 Species distributions

Traditionally, microbial species, including diatoms, are thought to be ubiquitous due to their small size, fast reproduction rates, high immigration rates and the ability to create resting spores or cells (Finlay, 2002; Finlay and Fenchel, 2004). According to this "theory of ubiquity", species distributions would be solely dictated by local environmental conditions as dispersal limitation would not exist. Although many diatom species may be cosmopolitan and thriving in all locations where local environmental conditions are favourable, a number of studies have reported endemism and restricted distributions among diatom species (Vanormelingen *et al*., 2008; Jüttner *et al.* 2010). This implies that, like documented for macroorganisms, small microbial taxa, such as diatoms, have biogeographical patterns (Martiny *et al.*, 2006; Astorga *et al*., 2012; Nemergut *et al*., 2013) related to historical factors and dispersal limitation (Vyverman *et al*., 2007; Verleyen *et al*., 2009).

Together with the large scale factors, diatom distributions are determined by the species-specific tolerances and preferences of environmental conditions. The ranges of tolerance of certain environmental variables, for instance the temperature and pH, have been determined for a number of diatom species (Weckström *et al*., 1997b; Mi-

chels *et al*., 2006; Andrén and Jarlman, 2008). As discussed earlier, species from a global species pool go through environmental filters operating at multiple scales to occur at a locality (see Fig. 1). The acting of such filters is supported by multiple studies observing restricted distributions of freshwater diatoms due to glaciation history and dispersal barriers, for example (Van de Vijver and Beyens, 1999; Vyverman *et al*., 2007; Vanormelingen *et al*., 2008). At a local scale, diatom distributions are affected by water chemistry and habitat characteristics, major ion concentrations being perhaps the most important variables for many species (Soininen, 2007). Diatom species distributions are most likely governed jointly by local environment and large scale factors, and the relative importance of these factors may depend on the spatial scale of observations (Soininen, 2007; Tang *et al.*, 2013). In spatially small data sets, climatic and geological gradients are small and thus local environmental factors seem to dictate species distributions (Martiny *et al.*, 2011). However as the geographical scale increases, the effect of dispersal limitation may override the influence of the local environment (Martiny *et al*., 2006).

1.2.3 Community composition

Diatom community composition is often described as the relative abundances of species in a site. The relative abundances of species in the community describe both the species tolerance and the preference towards the prevailing conditions and the competitive strength of the species as a species is most abundant in favourable conditions (Tilman, 1977; McCormick, 1996). Community composition can shift even if the species richness or the productivity remain unchanged (Hoagland *et al*., 1996). Therefore, it is perhaps the most relevant measure of the biological response to changes in the environment.

Factors determining stream diatom community composition have been widely studied. Communities are most often found to respond to water chemistry, most importantly conductivity, pH and nutrients (e.g., Soininen *et al*., 2004; Michels *et al*., 2006; Virtanen and Soininen, 2012), physical habitat characteristics (for example, current velocity, substratum size) (e.g., Passy, 2001; Michels *et al*., 2006; Jüttner *et al.* 2010), and stream degradation (e.g., Lavoie *et al*., 2006; Moravcova *et al*., 2013). However, community composition is not merely determined by the local environmental conditions, but also affected by catchment properties, such as land use, climate and spatial factors, that is, dispersal limitation (Weckström *et al*., 1997a; Leland and Porter 2000; Potapova and Charles, 2002; Soininen *et al.*, 2004; Urrea and Sabater 2009; Heino *et al*., 2010).

Changes in diatom community composition have been related to an increase in human land use and to the corresponding alterations in stream conditions (Walsh and Wepener, 2009; Chen *et al.*, 2016). The changed community is mainly composed of tolerant taxa and the sensitive taxa will recede. Distinct communities are found in streams under strong anthropogenic influence (Carpenter and Waite, 2000; Walker and Pan, 2006; Teittinen *et al.*, 2015), brownification of streams in catchments with wetlands (Pound *et al*., 2013), and in harsh low nutrient conditions, for example (Esposito *et al*., 2006). A growing evidence implies that diatom communities are often strongly spatially structured (Soininen *et al.*, 2004; Heino *et al.*, 2010; Liu *et al*., 2016). Spatial factors, i.e. position in geographic regions or stream system, and land use can sometimes explain more of the variation among diatom communities than do local environmental variables (Charles *et al*., 2006; Heino *et al*., 2010; Liu *et al*., 2016; Jyrkänkallio-Mikkola *et al*., 2017). For instance, Virtanen and Soininen (2012) found

that diatom community composition varied more than corresponding local environmental stream variables between geographical regions.

Additionally, physical and biological disturbances, such as the flow regime and grazing, alter community composition (Peterson, 1996; Steinman, 1996). For example, extreme storm events can detach a large part of the biofilm leaving only tightly attached low-profile species in the habitat (Lake, 2000; Schneck *et al*., 2017). Diatom communities have temporal fluctuations and succession, which re-continues from an earlier stage after a disturbance (Smucker and Vis, 2013).

1.3 Biomonitoring

"*Life is the ultimate monitor of environmental quality*", Lowe and Pan (1996).

Although some harmonization within larger geographical regions, for instance the EU, have been taking place, many countries have developed their own monitoring programmes and indices for water quality assessment because of the degradation of freshwater habitats (Whitton, 1991; Prygiel *et al*., 1997). Benthic diatoms have been widely used as biological indicators to assess the ecological status of freshwater ecosystems as they reflect the water quality over a longer period of time than a snapshot water chemistry sampling (Sandin and Verdonschot, 2006). Benthic diatoms are found to be suitable bioindicators as they are small and easy to sample, species-rich, they respond fast to changes in the environment due to their short life cycles, and they are sessile in their habitat, thus reflecting well the prevailing conditions (Lowe and Pan, 1996; Smol and Stoermer, 2010). In addition, the environmental preferences and tolerances are known for many taxa (e.g., van Dam *et al.* 1994, but see Potapova and Charles, 2007). A good indicator species possesses a narrow range of tolerance towards an environmental variable. Similarly, the ability of diatoms to indicate environmental variation has been widely utilized in palaeolimnology, where the past environmental and climatic conditions are inferred from diatom communities derived from lake sediments (Smol, 2010).

Recently, studies have implied that diatom indices may have a lower predictive power in regions other than those they were created in (Potapova and Charles, 2007; Besse-Lototskaya *et al.*, 2011). This may be due to local adaptation or lack of shared species between regions. Additionally, species identified at species-level may also include subspecies, which vary in their responses towards local environmental conditions (Round, 2004; Vanormelingen *et al*., 2008; Rose and Cox, 2014) or species distributions may be constrained by large scale factors such as climate or geology (Weilhoefer and Pan, 2006; Jüttner *et al*., 2010). This arises the question whether the responses of individual species or even communities are somewhat context dependent, i.e. the main determinants of species distribution and abundance vary between environments and geographic locations. Furthermore, this arises a concern of the reliability of bioindicators especially in changing climate and other environmental conditions.

1.4 The study aims

Although the effects of environmental variables on diatom species have been widely examined, more knowledge about species distributions and abundance and the complex interactions between the factors affecting them, is needed. This would help to predict the effects of changing climate and anthropogenic stressors on these pivotal primary producers and bioindicators in streams. This study will address the biogeography of diatoms and especially the arisen concerns of the reliability of diatoms as bioin-

dicators in a changing climate.

Thus, the aims of this thesis are:

- To investigate the relative importance of local environmental and climatic factors on diatom species distributions at a regional scale (c. 1200 km) in Finland. This would reveal whether species distributions are mainly determined by local environmental variables or do climatic variables also explain their distributional patterns (Paper **I**).

- To analyse whether the relative importance of factors governing diatom species distributions differs between human impacted and pristine stream sites. Such a study would reveal whether the most important determinants of diatom species occurrence are context dependent (Paper **II**).

- To explore the direct and indirect effects of hierarchical factors, i.e. climate, land use and water chemistry, on diatom species richness, beta diversity and community composition. This would reveal the effect of the most relevant factors on stream diatoms (Paper **III**).

- To investigate the ability of stream diatom communities to predict local environmental and climatic conditions using multiple modelling techniques. This would reveal whether diatom communities are reliable bioindicators for water chemistry and climate, and also, whether their predictive ability varies between modelling methods (Paper **IV**).

These investigations are performed using novel modelling methods and approaches, for example machine learning techniques such as boosted regression tree (BRT) and random forest (RF) (Elith *et al*., 2006; Cutler *et al*., 2007; De'ath, 2007), structural equation models (piecewise SEM; Lefcheck, 2016) and local contribution to beta diversity (LCBD; Legendre and De Cáceres, 2013), but also with more traditional methods, for instance redundancy analysis (RDA) and weighted averaging (WA), widely used in diatom studies (e.g., Leland and Porter, 2000; Soininen and Niemelä, 2002; Tudesque *et al.*, 2014; Liu *et al*., 2016).

2 Material and methods

2.1 Study area and sites

The study was conducted in Finland, northern Europe at a regional scale extending c. 1200 km (60° – 70° N, 20° – 32° E) (Fig. 3). A comprehensive data set (n = 392) of stream diatom assemblages and stream variables was gathered from three existing data sets and supplementary data collected during the study. The full data set comprised 56 stream sites collected from central Finland in 1986 (Eloranta, 1995), 141 sites collected between 1996 and 2001 encompassing a wide latitudinal gradient (Soininen *et al*., 2004), 30 sites collected in 2004 from northern Finland, and additionally, 105 sites collected in 2014 from western Finland (Jyrkänkallio-Mikkola *et al*., 2017) and 60 sites collected between 2014 and 2015 from southern, eastern and northern Finland. Thus, the data set covered wide gradients of both local environmental, catchment and climatic variables. Most of the stream sites (n = 305) were independent, i.e. having fully separate catchment area from the other stream sites, yet 87 of the sites had nested catchments, i.e. they were located downstream from some other stream sites in the study.

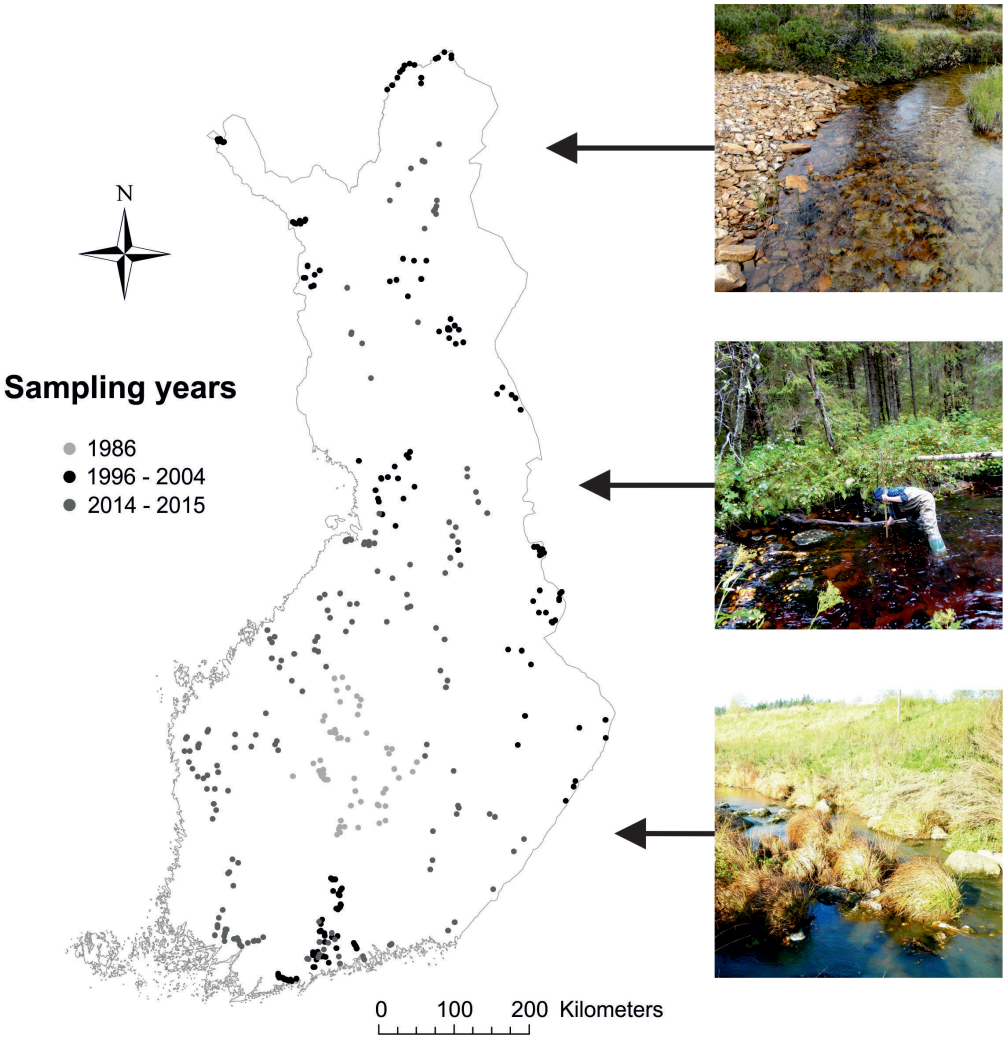


Figure 3 A map of the study area in Finland (60° – 70° N, 20° – 32° E) comprising 392 stream sites. The sites in the map are categorized by the sampling years. The pictures show examples of typical stream types in each region. Photo credits: V. Pajunen and J. Jyrkänkallio-Mikkola.

2.2 Sampling and analyses

All samples were collected during the base flow conditions in July to September with harmonized sampling procedures. From each sampling site, diatom samples were obtained by brushing five to ten approximately cobble sized stones collected along the reach (c. 10 m). Water samples were collected concurrently, but in a few sites water chemistry data were obtained later from the national water quality database. Water samples were analysed for total phosphorus (TP), pH, conductivity and water colour. In the field, physical properties, i.e. current velocity, canopy shading, stream width and depth, were measured.

Diatom samples were prepared by cleaning them from organic material using wet combustion with acid or hydrogen peroxide and preserved in Naphrax or Dirax. Diatoms were identified to the lowest possible taxonomic level ac-

cording to Krammer and Lange-Bertalot (1986 – 1991) and Lange-Bertalot and Metzeltin (1996) and enumerated (250 – 500 frustules per sample) using phase contrast light microscopy (magnification 1000×). See further details of sampling and sample analyses in papers **I-IV**.

2.3 Catchment and climatic data

The catchment area was defined for each stream site by calculating flow direction and accumulation patterns from a digital elevation model (grid resolution 10×10 m, National Land Survey of Finland, 2013) to each sampling point. The relative percentages of different land use classes were determined for each catchment (CORINE Land Cover data, 20×20 m, Finnish Environment Institute, 2013). Land use classes of artificial and agricultural areas were combined as an anthropogenic land use class. The data sets were classified in papers **II** and **III** based on the amount of anthropogenic land use into two groups: human impacted sites (> 5% of anthropogenic land use) and pristine or reference sites (< 5% of anthropogenic land use). The catchments and the land use variables were determined using ArcGIS 10.3.1 software. For more detailed description of the catchment data, see papers **II** and **III**.

Three climatic variables were chosen for the models: growing degree days (GDD), precipitation sum from May to September (PRECS) and water balance (WAB). GDD represents the thermal regime and the length of the growing season (adjusted to 5 °C). PRECS and WAB represent the water input and availability in the stream system. The values of the climatic variables for each sampling site were obtained from a 10×10 km resolution climatic grid, which covered the years 1981 – 2000 (Finnish Meteorological Institute; Venäläinen and Heikinheimo, 2002). For more information on the climatic data, see papers **I-IV**.

2.4 Statistical analyses and modelling

2.4.1 An overview of the statistical analyses

All the statistical analyses and modelling were performed in R software (versions 3.1.1 – 3.3.3; R Development Core Team, 2016). The climatic and environmental variables were tested for collinearity with the nonparametric Spearman's rank correlation coefficient and the collinearity was relatively low in all data sets ($r_s < |0.80|$). The interrelations between climatic and environmental variables and the relationship between the environmental variables and diatom assemblages were examined by performing principal component analysis (PCA), RDA and partial redundancy analysis (pRDA) using the package "vegan" (Oksanen *et al*., 2015) (**I** and **IV)**. A variable representing the variation among community compositions was created by performing non-metric multidimensional scaling (NMDS) (in "vegan") for diatom assemblage data in paper **III**. The values derived from the first axis of NMDS for each site indicate the variation among community composition between the sites (Hough-Snee *et al.*, 2014). LCBD, the variable representing the community uniqueness and the contribution to regional beta diversity, was calculated using the function "beta.div" according to Legendre and De Cáceres (2013) (**III**).

2.4.2 Species distribution models

Species distribution models (SDMs) were performed for diatom species collected from 277 sites between the years 1986 and 2004 (**I**) and from human impacted stream sites (n = 164) and pristine stream sites (n = 164) collected between the years 1986 and 2015 (**II**). These datasets comprised presence-absence data of diatom

species occurring in at least 5% and the maximum of 95% of the sites in each dataset. Collectively, in paper **I**, SDMs were conducted for 157 taxa, and in paper **II**, for 110 taxa occurring in both human impacted and pristine sites. In paper **I**, three sets of SDMs were performed for each species: environment-only, climate-only and full models. In paper **II**, climate and full models were performed for the species occurring at both human impacted and pristine sites. Environment-only models consisted of three local environmental variables (TP, conductivity and water colour) and all climatic models included three climatic predictors (GDD, PRECS and WAB). The full models in paper **I** combined the variables from the environment-only and the climate-only models, whereas the full models in paper **II** comprised the three climatic variables in addition to TP, conductivity, pH, water colour, shading and current velocity.

The SDMs were conducted using the BIOMOD (Thuiller *et al*., 2009) (**I**) or BIOMOD2 framework (Thuiller *et al*., 2016) (**II**). In paper **I**, potential differences in model performances associated to the methodologies were considered by using four different modelling techniques: generalized linear model (GLM), generalized additive model (GAM), BRT and RF. The SDMs in paper **II** were performed using BRT. The main principles of the modelling methods and references for further information are listed in Table 1. The model performances were evaluated with a cross validation approach: SMDs were fitted four times by evaluating a random sample of 70% of the data against the remaining 30%. The model performances were determined from the validation data set by calculating the area under the curve of a receiver operating characteristics plot (AUC; Fielding and Bell, 1997) and true skill statistics (TSS; Allouche *et al*., 2006). In paper **I**, the differences among the predictive performances of environment-only, climate-only and full models were tested with the non-parametric Wilcoxon signed rank test. Finally, the relative importance of each variable for each species was calculated according to Thuiller *et al*. (2009). See more details of the model fitting and evaluation in papers **I** and **II**.

2.4.3 Structural equation models

In paper **III**, SEMs were used to investigate the links among climatic, catchment and local environmental variables, and their effect on diatom diversity (alpha and beta) and communities. The data set comprised 143 stream sites collected between the years 1986 and 2004. All the sites represented individual streams, i.e. the data did not include any nested catchments. The models were conducted using the piecewiseSEM package (Lefcheck, 2016). Two predictor variables were chosen from each spatial scale: climatic (GDD and PRECS), catchment (anthropogenic land use and wetlands) and local environmental (TP and conductivity). The climatic variables were set as exogenic variables and the catchment and local environmental variables as endogenic variables. SEMs were conducted separately for species richness, community composition (the first axis of NMDS) and local contribution to beta diversity (LCBD).

The models were built by including all the potential causal links between the variables. The non-significant linkages were gradually removed maintaining the causal structures of the models. The composite variables were composed using the second polynomial terms of GDD and PRECS to account for non-linear relationships between GDD and the catchment variables, GDD and community composition, and PRECS and conductivity. Spatial autocorrelation was accounted for by using the spatialCorrect function. The criterion of model parsimony and goodness of fit (Fisher's C, $P > 0.005$) was used to assess

Table 1 The main principles of the used modelling methods and references to further information and relevant research.

Method	Main principles	Reference	Applied in	Paper
BRT	A machine learning technique, in which a boosting method is used to minimize e.g. deviation in predictions by growing gradually a sequence of simple regression trees.	*Friedman et al., 2000* *De'ath, 2007* *Elith et al., 2008*	Lake plankton groups, SDMs *Soininen et al., 2013* Stream diatoms and multiscale factors *Jyrkänkallio-Mikkola et al., 2017*	**I** **II** **IV**
GAM	Nonparametric extensions of GLMs, which use smoothers when estimating the form of relationship between predictors and response variables.	*Hastie and Tibshirani, 1990*	Lake plankton groups, SDMs *Soininen et al., 2013*	**I**
GLM	Mathematical extensions of linear models, which allow non-linearity and non-constant data variance.	*McCullagh and Nelder, 1989*	Lake plankton groups, SDMs *Soininen et al., 2013*	**I**
MAT	A technique, in which an analogue is compared numerically to species abundance data using a measure of dissimilarity.	*Overpeck et al., 1985*	Lake diatoms and pH *Battarbee et al., 2005*	**IV**
piecewise SEM	A version of SEM, in which local estimation is used to individually evaluate linkages between variables structured as a set of separate linear equations.	*Lefcheck, 2016*	Plant ecology and climate *Jing et al., 2015*	**III**
RF	A machine learning technique, in which a forest of regression trees is grown with a randomized subset of predictors to produce accurate predictions without overfitting the data.	*Breiman, 2001*	Stream diatoms and anthropogenic stressors *Larras et al., 2017*	**I** **IV**
WA	A technique, which computes weighted average for each species and is based on the assumption of unimodal relationships between variables.	*ter Braak and van Dam, 1989*	Lake diatoms and pH, temperature *Weckström et al., 1997b* Stream diatoms and TP *Soininen and Niemelä, 2002*	**IV**
WA-PLS	An extension of WA, which involves a weighted inverse deshrinking regression.	*ter Braak and Juggins, 1993*	Stream diatoms and TP *Potapova et al., 2004*	**IV**

whether the model could be accepted. A more detailed description of the modelling process can be found in paper **III**.

2.4.4 Inference models

In paper **IV**, climatic and local environmental variables were inferred from diatom assemblage data (214 taxa in total) of 227 stream sites collected between the years 1986 and 2004. Five modelling methods were used as calibration and inference tools: WA, weighted averaging partial least squares (WA-PLS), modern-analogue technique (MAT), BRT and RF. This allowed a comparison of the model performance among

the modelling techniques. The relative abundances of diatom species were used as predictors in all models and the response variables included GDD, PRECS, WAB, conductivity, TP and water colour. The models conducted using WA, WA-PLS and MAT were fitted using the RIOJA package (version 0.8-5; Juggins, 2013), MAT using ANALOGUE package (version 0.10-0; Simpson and Oksanen, 2013), BRT using GBM package (version 1.6-3.1; Ridgeway, 2010) and RF using randomForest package. The performances of all models were assessed using leave-one-out cross-validation and estimated using the root-mean-square error of prediction (RMSEP) and the coefficient of determination (r^2). The relative importance of the predictor variables, i.e. diatom species, was estimated in the BRT models according to Friedman (2001). Spatial autocorrelation was assessed for each climatic and environmental variables using pgirmess package to create spatial correlograms. This demonstrated that in the model residuals the spatial autocorrelation was considerably smaller than in the raw data, and hence, the uncertainty of model estimation was reduced. A more detailed description of the methodology, calibration and model fitting can be found in paper **IV**.

3 Results and discussion

3.1 Papers I and II: Drivers of species distributions

This study revealed a fundamental effect of climate on stream diatom distributions (Fig. 4). A climatic variable had the greatest relative importance in SDMs consisting of both local environmental and climatic predictors: GDD in the full data set (227 sites) (**I**) and WAB in both pristine and human impacted sites (**II**). Also, PRECS had a significant influence in the SDMs. This corresponds to the previous research indicating that also microbial species can follow similar biogeographical patterns related to climate to those observed among macro-organisms (Weckström *et al*., 1997a; Vyverman *et al*., 2007; Verleyen *et al*., 2009, Berthon *et al*., 2014). The results of this study suggest that, in addition to the local environmental filtering, the distributions of stream diatoms are constrained by large scale environmental filters, such as climate.

Climatic factors, here GDD, WAB and PRECS, set the limits to diatom species distributions as the thermal regime directly affects species (Brown *et al*., 2004). Diatom species have an optimum temperature for growth and some species have narrow temperature ranges (Patrick, 1971; Weckström *et al*., 1997b). But also, the effects of climate are manifested through intermediate variables, such as productivity, land use and the flow regime (Stevenson, 1997). A diatom species can prefer a certain flow regime as species vary in their tolerances towards flow velocities (Passy, 2001). Some species, for example *Achnanthes pusilla* (currently regarded as *Rossithidium pusillum*), thrive in harsh conditions with cold temperature, low organic matter and nutrient content, conditions that are greatly driven by climate (**IV**). The indirect effects of WAB and PRECS are connected to catchment properties as they amplify the impact of land use (Andersen *et al*., 2006; Jeppesen *et al*., 2009; Arvola *et al*., 2015). Precipitation empowers the flow of essential materials, such as nutrients, from atmosphere (N_2) and land to streams and subsequently to benthic diatoms (Moss, 1998; Allan and Castillo, 2007; Kryza *et al*., 2012).

Conductivity was the most influential local environmental variable and it has been recognised as an important factor for benthic diatoms in boreal streams (Soininen *et al.*, 2004). It had the second greatest relative importance in hu-

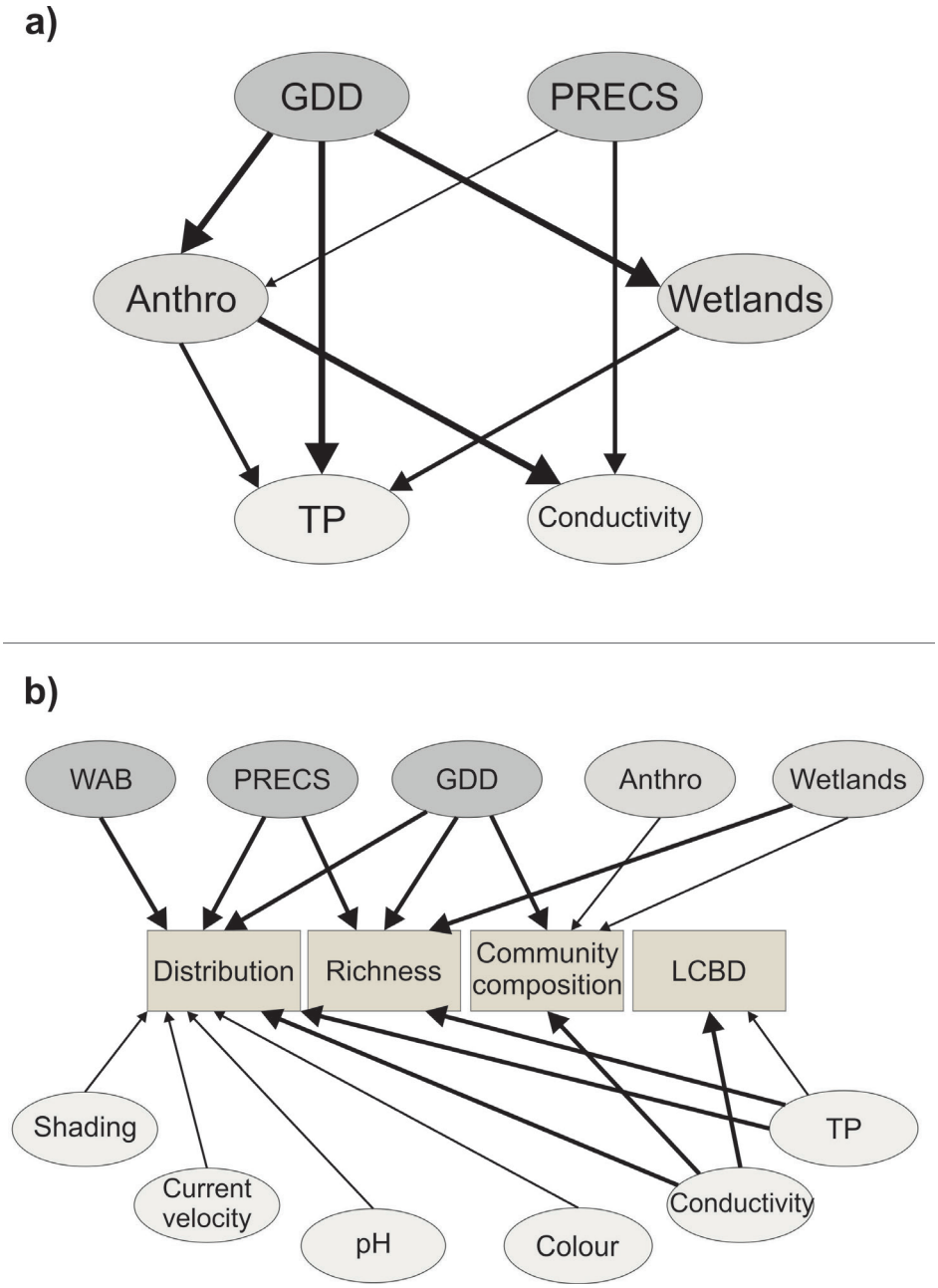


Figure 4 A conceptual model of the factors influencing stream diatoms based on this thesis. a) The potential causal links between the most important climatic, land use and water chemistry variables based on the results of SEMs (**III**). The arrows are scaled to match the magnitude of the effect. Only significant paths are shown ($P < 0.05$). b) A summary of the factors affecting stream diatoms. The thicker arrows represent greater relative importance. Richness, community composition and LCBD were modelled using only the factors presented in the figure a (**III**). The distributions were modelled using climatic and local environmental factors (**I** and **II**).

man impacted sites (**II**) and in the full data set (**I**), which covered a wide anthropogenic gradient. Under anthropogenic influence, the local environmental variables had greater overall effect

than climate (**II**). This most likely results from the wider gradients of water chemistry variables related to anthropogenic land use, including for instance high nutrient concentrations and conductivity (Fig. 5). As a consequence, water chemistry does not reflect much the large scale climatic conditions and geology, but is rather an imprint of human actions (Allan, 2004; Wang *et al*., 2008; Rothwell *et al*., 2010). Except for conductivity, the other measured local environmental variables did not reach high relative importance on species distributions on average. Nevertheless, the distributions of individual taxa had highly variable responses towards different local environmental and climatic variables highlighting the species-specific niche requirements among diatoms (e.g., van Dam *et al*., 1994).

The model performance increased from environmental-only to climate-only models and was the greatest in the full models (**I**). The environment-only model predicted false occurrences for some species, that is, the model assumed that species would occur at a certain site based on the stream water chemistry. This result suggests that stream diatoms may experience dispersal limitation in the study area or the sampling did not detect the species at a site (Heino *et al*., 2010; Ashcroft *et al.*, 2017). The performance of the full models, including both climatic and local environmental variables, was not higher than the performance of climatic models either in human impacted or pristine sites (**II**). This indicates that climatic variables, which have both an immediate and indirect impact on diatoms and abiotic

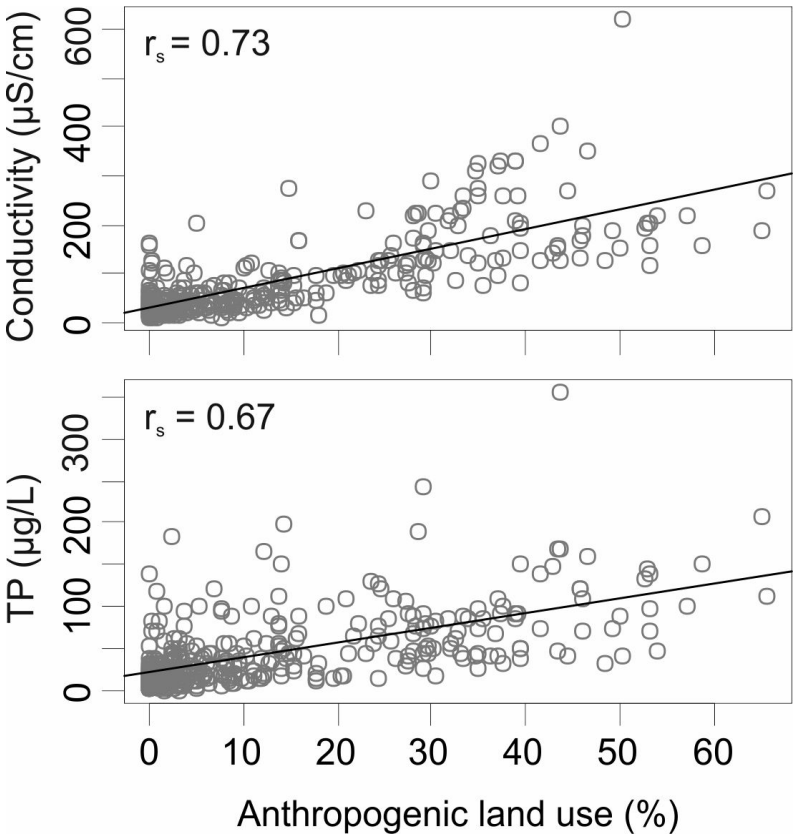


Figure 5 The measured conductivity and TP concentrations in relation to the percentage of anthropogenic land use in the stream catchments (n = 380). Correlations were calculated with Spearman's correlation coefficient (r_s).

factors in streams, are able to predict overall species distributions more accurately than a certain water chemistry variable or even a set of physicochemical variables. Based on existing literature, a low predictive performance for the SDMs was expected due to the stochastic distributions of microbes (Soininen *et al*., 2013). On the contrary, the majority of SDMs in papers **I** and **II** had at least intermediate performance.

Notably, the relative importance of climatic and environmental factors varied between human impacted and pristine stream sites, and even the most important factor for individual species differed between the site groups (**II**). This indicates that there is a clear context dependency in the factors driving diatom species distributions. For example, GDD was a significantly more important driver of species distributions in pristine than in human impacted streams. It suggests that in pristine streams, the local environmental conditions are mainly dictated by large scale climatic factors and geology. Whereas under human influence, the effect of anthropogenic land use may override the natural hierarchy of multiscale factors governing the stream conditions. Context dependency may challenge the predictions of diatom species distributions as the relative importance of the main drivers could depend on site characteristics and vary among geographic regions (Charles *et al*., 2006; Jüttner *et al*., 2010).

3.2 Paper III: Drivers and patterns of diversity

Paper **III** demonstrated a strong link between factors operating at multiple spatial scales affecting stream diatoms (Fig. 4). GDD had a strong positive impact on the relative amount of anthropogenic land use and wetlands, which in turn influenced water chemistry: TP and conductivity. Nutrient availability correlated positively with increased diatom richness, which corresponds to previous findings (Liess *et al*., 2008; Passy, 2009; 2010). Although GDD had an indirect positive effect on species richness through anthropogenic land use and nutrients, the SEM method was able to separate a negative direct effect of GDD on richness not related to other predictors in the model. This result implies that diatoms may be less prone to interspecific competition or grazing in colder and perhaps more harsh conditions (Liess *et al*., 2008; Piggott *et al.*, 2015). Diatoms tolerate colder water temperatures than other benthic algae (green algae and cyanobacteria) (Gudmundsdottir *et al.*, 2011) and are often the most species-rich primary producers in cold mountain streams (Hieber *et al*., 2001; Rott *et al*., 2006). Soininen *et al*. (2016) found a similar pattern of diatom species richness increasing with latitude in a global study, and proposed that the pattern may be linked to the patterns of catchment properties influencing water pH, for instance. Additionally, in northern tundra regions where GDD is low, stream light levels are high due to the low amount of terrestrial and aquatic vegetation (see Fig. 3). High light availability in the stream bottom could result in high diatom species richness (Liess *et al*., 2008). Whether the negative correlation between GDD and species richness observed in this study originate from climate, catchment properties, local stream conditions or biotic interactions, needs further investigation.

PRECS and TP had positive effects on diatom species richness (**III**). The effect of PRECS most probably reflects the impact of some unmeasured variables, such as micronutrients, whose concentrations and bioavailability are strongly regulated by weathering, acidification and other processes related to precipitation (Rothwell *et al*., 2010; Kryza *et al*., 2012). The amount of wetlands affected diatom richness negatively, unlike in a study conducted in hard water streams in the continental U.S.A. (Passy, 2010). In contrast

to temperate wetlands in the U.S.A, streams affected by boreal peatlands are more acidic and have lower light levels due to humic substances in stream water. This can create a rather hostile environment for many aquatic species (Hall *et al*, 1980; Guerold *et al*., 2000). However, Pound *et al*. (2013) reported an increase in diatom richness related to the amount of dissolved organic carbon in acidic streams affected by wetlands in the U.S.A. This further highlights the complexity and a possible context dependency in the factors affecting diatom species richness.

The local contribution to beta diversity was highest in the southernmost and the northernmost sites in Finland. The LCBD values were inversely correlated with species richness. This indicates that sites with high LCBD values harbour a low number of unique species which are well-adapted to prevailing stream conditions and are able to outcompete other species (Legendre and De Cáceres, 2013; Heino *et al*., 2017; Vilmi *et al*., 2017). Also, the stream conditions may be unfavourable for most species, which hinders colonization: harsh, low-nutrient conditions in the north and strongly human impacted and maybe degraded sites with high conductivity in the south. This agrees with the findings of Smucker and Vis (2013) who reported that communities in extreme environments were composed of few adapted species. The large scale climatic and catchment scale variables influenced LCBD only indirectly through TP and conductivity. LCBD was strongly affected by conductivity. This corresponds to the fact that conductivity was the most influential local environmental factor driving diatom species distributions (**I** and **II**).

The SEM could explain only a small fraction of diatom species richness ($r^2 = 0.13$). This could be expected as the variation in diatom species richness is affected by a wide variety of factors with complex interactions, and thus, it cannot be fully explained by just a few factors. Additionally, the data used in this study comprised sites from different stream orders. As richness can be higher in larger streams (Stenger-Kovacs *et al*., 2014) or peak in the mid-order streams (Vannoute *et al.*, 1980), this is likely to increase the unexplained variation in the models.

3.3 Paper III: Drivers of community composition

In the SEM, community composition showed shifts along the conductivity gradient and, on the other hand, along the amount of wetlands (**III**). Among the other explanatory variables in the model, conductivity, GDD and wetlands were the key factors affecting communities in the data set (Fig. 4). High conductivity seems to create conditions tolerated by specialised taxa. This was confirmed as conductivity was also the main driver of LCBD. Streams influenced by wetlands, i.e. boreal peatlands, are typically occupied by species from the genus *Eunotia*, which have a high tolerance for low pH and low light conditions (van Dam *et al*., 1994). Anthropogenic land use also had a notable influence on community composition through conductivity, indicating that high conductivity is mainly originated from human impacts in this study (Fig. 5). With this in mind, the effect of conductivity may reflect other variables from anthropogenic sources, which sets limits to species occurrences and abundances: for instance, toxicants (Rai *et al*., 1981).

3.4 Papers I-IV: The effect of scale and human impact

The study scale is a major factor influencing the results gained from the observational studies of freshwater diatoms (Soininen, 2004). Traditionally, the ecological niche requirements of diatom species have been investigated in regional studies (for example, van Dam *et al*., 1994; Fore

and Grafe, 2002). However, a growing evidence suggests that these observed species responses may not correlate with those observed in other regions indicating that diatoms are influenced by spatial factors or are locally adapted (Potapova and Charles, 2007; Chen *et al*., 2016). It has been proposed that at local (< 100 km) and regional scales (100 – 3000 km), the importance of local environmental factors to microbial distributions may be more profound, and as the spatial scale increases to continental and global scale, large scale factors, such as history, climate and geology, may become more important (Martiny *et al.*, 2006; Astorga *et al*., 2012). This seems evident as large scale factors may not vary enough at smaller, regional scales where the climatic or geological gradient is narrow (Martiny *et al.*, 2011). Thus, one may not see forest for the trees as the environmental heterogeneity derived from climate and geology is only detected from sufficient distance or spatial scale. However, in this study, conducted at a regional scale (c. 1000 km), the effect of climate on stream diatoms was clear and sometimes even stronger than those of the local environmental variables (**I-IV**). The study scale may also affect the observed shape of the species' response. In nature, species' responses to the contemporary conditions are seldom linear (Mittelbach, 2007). Therefore, to observe unimodal or yet more complex responses for species, sufficiently wide gradients of environmental variables need to be considered.

The effect of human impact on both abiotic stream conditions and diatoms was evident in this study (**II** and **III**). The most influential local environmental variable, conductivity, and in some extent TP, were significantly related to anthropogenic land use (Fig. 5). The high values of LCBD and the shift in community composition related to conductivity indicate that the diatom communities in human impacted streams consist of unique species tolerant to pollutants from anthropogenic sources (Lavoie *et al*., 2006; Moravcova *et al.*, 2013) (**III**). However, nutrients originated from anthropogenic land use led to a higher diatom species richness, a pattern observed also by Johnson and Angeler (2014) and Jyrkänkallio-Mikkola *et al.* (2017), to name a few. In paper **II**, the local environmental factors were more important than the large scale climatic factors in human impacted streams. This suggests that anthropogenic land use has somewhat universal effect on stream conditions, and thus, the local environmental variables have a similar effect on diatoms among human impacted streams across spatial scales.

3.5 Paper IV: Diatoms as environmental indicators

Diatom assemblages turned out to be reliable indicators of both climatic and local environmental factors with all five modelling methods used (**IV**). Noteworthy, climatic variables were predicted more accurately than local environmental variables by diatom assemblages. This suggests that diatom assemblages respond to a number of proximate variables, such as the temperature and flow, which are driven by climate (Stevenson, 1997). Thus, diatom assemblages could serve as useful proxies for certain climatic conditions and could be used for monitoring the effects of climate change. However, such proxies should be used cautiously as diatoms could be affected also by other large scale factors such as geology and history. A number of studies have reported that diatom indices are specific to the region of their origin, and therefore specific indices should probably be constructed for each geographic region (Charles *et al*., 2006; Potapova and Charles, 2007; Bottin *et al.*, 2014). As found in paper **III**, the main drivers of species distributions may be context dependent, even for a single species. For example, the relative importance of

GDD on species distributions was significantly lower in human impacted than in pristine sites. Therefore, also the assemblage responses may change among geographic regions or along an anthropogenic gradient. Thus, it is suggested that when diatoms are used as proxies indicating (paleo)climatic conditions, the potential influence of geology, evolutionary history and human impact on the calibration data should be accounted for, especially when using proxy data from another geographical region.

The machine learning techniques, BRT and RF, were robust modelling methods for predictive purposes and in finding species with a good indication value (**IV**). One of the advances of these methods is that they are able to take into account complex nonlinear responses (Cutler *et al.*, 2007; Elith *et al*., 2008). BRT was able to detect thresholds for species along climatic and environmental variables corresponding to earlier findings in freshwaters (Potapova *et al*., 2004; Soininen *et al.*, 2013). The method traditionally used in predictive modelling, WA, may be too simplistic as it assumes that species have unimodal responses towards environmental variables (ter Braak and van Dam, 1989). Here, WA was outcompeted in predictive performance by the other methods except in TP predictions. This may imply that diatom species indeed have unimodal responses towards TP (Soininen and Niemelä, 2002). The indicator species for local environmental variables identified in this study corresponded well to previous studies (e.g., Fore and Grafe, 2002; Potapova and Charles, 2003; Rimet *et al.*, 2005; Urrea and Sabater, 2009). A good climatic indicator species may respond to varying effects of climatic variable, thus reflecting a certain type of environment which is strongly governed by climate. For example, *Achnanthes pusilla* (i.e. *Rossithidium pusillum*), a good indicator of GDD, reflects harsh cold environments with low productivity.

4 Conclusions and future aspects

4.1 Microbial world in a changing climate

Despite their small sizes, fast life-cycles and dispersal rates, aquatic microbes, such as stream diatoms, are not safe from the effects of climatic changes and anthropogenic stressors. As this study revealed, the microbial world is driven not only by local physicochemical variables, but also by large scale factors such as climate. Although climatic factors operate mostly indirectly via a myriad of other variables, they may also have immediate influences such as thermal conditions and disturbance by storm events. Thus, energy and water are undoubtedly the ultimate driving forces of many stream diatoms as they are for numerous other organisms (Hawkins *et al.*, 2003).

The relative importance of climatic and local environmental factors vary among individual species. The occurrences of some species are more related to climatic conditions and others to certain stream physicochemical variables, such as conductivity or water pH. Also, catchment properties, such as anthropogenic land use and wetlands, can be reflected in diatom community composition through their effect on stream conditions. This variety of environmental preferences may in fact be the reason for the high species richness of benthic diatoms. This study found that some sites located in the northernmost Finland harboured distinct diatom communities comprising a low number of species with high contribution to regional beta diversity. As species-rich communities are better buffered against environmental change (Chapin *et al.*, 1997), these stream communities may be endangered. Similarly, the communities in human impacted streams may comprise only a few taxa which are pollutant-

tolerant and not found in more pristine streams. Notably, the importance of environmental factors may vary along an anthropogenic land use gradient thus highlighting the impact of human actions on stream biota.

The ongoing climatic and other environmental changes will evidently affect diatom species distributions. Warming temperatures and increased precipitation in boreal regions in concert with increasing anthropogenic land use will undoubtedly enhance productivity. This may increase interspecific competition and grazing pressure, both of which have notable effect on diatom occurrences and abundances. Extreme storm events are projected to increase the frequency and magnitude of disturbances caused by the flow. As assemblages are more frequently damaged by high currents, low-profile species, fast colonizers and cosmopolitan species with a high dispersal ability may be favoured. Specialist species, adapted to harsh cold conditions and having narrow thermal ranges or poor competitive ability in more productive environments, are thus prone to be endangered. The climatic and other environmental change produce novel stream conditions and therefore the composition of diatom communities is yet unforeseen. Based on the context dependency found in this study, it may be expected that the species' responses may differ from the present responses in novel conditions. This calls for more knowledge of the mechanisms behind the observed context dependency in species responses.

The future diatom studies should involve progress and be open-minded to new applications. The proxy variables, such as climate and land use, contain the effect of a variety of proximate variables, which may be difficult or costly to measure. Therefore, climatic and land use variables can act as important tools in predicting future species distributions and shifts in community composition. Also, the development of more efficient modelling algorithms in the near future and the usage of super-computers could enable the addition of more predictors in the models without compromising the model robustness. This would allow a more in-depth investigation of the complex drivers of stream diatom occurrences and abundances. For example, the biotic interactions among diatoms and other organisms involve an enormous number of linkages between species and among trophic levels. Open data sources would enable the usage of large data sets covering large spatial scales. Finally, the usage of DNA sequencing methods, for instance metabarcoding, in species identification and enumeration in addition to traditional morphological identification would give new insights into diatom distributions and ecology (Zimmermann *et al*., 2015; Rimet *et al*., 2016).

4.2 Considerations for biomonitoring

From an applied perspective, there is a need for robust diatom indices for monitoring. This study presented that climate has a notable impact on stream diatoms. Furthermore, the relative importance of climate and local environmental factors can be context dependent. Therefore, it is suggested that the following statements could be accounted for in biomonitoring practises:

- Indices which have been developed elsewhere, may not reflect the focal environmental conditions reliably as species-specific and community-level responses may vary among geographic regions, climatic conditions and along an anthropogenic gradient. As a corollary, indices should be developed within the same geographical region where they are applied (Potapova and Charles, 2007).

- The most important drivers of diatom species occurrences can vary between human

impacted and pristine sites. In human impacted sites, the species occurrences may reflect the water quality well, yet in more pristine sites, climate can have a stronger effect. As a corollary, indices should be developed separately for human impacted and pristine sites.

- As climatic and other environmental changes may generate novel conditions, indices should be updated regularly. This would ensure that the species' responses towards environmental conditions are constantly up to date.

- The ability of diatoms to indicate climatic conditions may be utilized when monitoring the current effects of climate change on freshwater ecosystems.

- The new machine learning techniques, such as BRT and RF, are able to recognize complex interactions and response shapes, and thus, they are robust methods in developing new monitoring methods. These methods have many advantages, for example, they are very flexible in fitting various data.

References

Allan, J. D. (2004) Landscapes and riverscapes: the influence of land use on stream ecosystems. Annual Review of Ecology and Systematics 35, 257-284.

Allan, J. D. and Castillo, M. M. (2007) Stream Ecology: Structure and Function on Running Waters. 2nd edn. Springer, Dordrecht.

Allouche, O., Tsoar, A. and Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223-1232.

Andersen, H. E., Kronvang, B., Larsen, S. E., Hoffmann, C. C., Jensen, T. S., and Rasmussen, E. K. (2006) Climate-change impacts on hydrology and nutrients in a Danish lowland river basin. Science of the Total Environment 365, 223-237.

Andrén, C. and Jarlman, A. (2008) Benthic diatoms as indicators of acidity in streams. Fundamental and Applied Limnology 173, 237-253.

Arvola, L., Einola, E. and Järvinen, M. (2015) Landscape properties and precipitation as determinants for high summer nitrogen load from boreal catchments. Landscape Ecology 30, 429-442.

Astorga, A., Oksanen, J., Luoto, M., Soininen, J. Virtanen, R. and Muotka, T. (2012) Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules? Global Ecology and Biogeography 21, 365-375.

Baron, J. S., Poff, N. L., Angermeier, P. L., Dahm, C. N., Gleick, P. H., Hairston, N. G. Jr., Jackson, R. B., Johnston, C.A., Richter, B. D. and Steinman, A. D. (2002) Meeting ecological and societal needs for freshwater. Ecological Applications 12, 1247-1260.

Battarbee, R. W., Monteith, D. T., Juggins, S., Evans, C. D., Jenkins, A. and Simpson, G. L. (2005) Reconstruction pre-acidification pH for an acidified Scottish loch: a comparison of palaeolimnological and modelling approaches. Environmental Pollution 137, 135-149.

Battin, T. J., Sloan, W. T., Kjellberg, S., Daims, H., Head, I. M., Curtis, T. P. and Eberl, L. (2007) Microbial landscapes: new paths to biofilm research. Nature Reviews Microbiology 5, 76-81.

Berthon, V., Alric, B., Rimet, F. and Perga, M.-E. (2014) Sensitivity and responses of diatoms to climate warming in lakes heavily influenced by humans. Freshwater Biology 59, 1755-1767.

Besemer, K. (2015) Biodiversity, community structure and function of biofilms in stream ecosystems. Research in Microbiology 166, 774-781.

Besse-Lototskaya A., Verdonschot P. F. M., Coste M. and van de Vijver B. (2011) Evaluation of European diatom trophic indices. Ecological Indicators 11, 456-467.

Biggs, B. J. F. (1996) Patterns in benthic algae of streams. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 31-56). Elsevier, San Diego.

Bottin M., Soininen J., Ferrol M. and Tison-Rosebery J. (2014) Do spatial patterns of benthic diatom assemblages vary across regions and years? Freshwater Science 33, 402-416.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. and West, G. B. (2004) Toward a metabolic theory of ecology. Ecology 85, 1771-1789.

Buck, O., Niyogi, D. K. and Townsend, C. R. (2004) Scale-dependence of land use effects on water quality of streams in agricultural catchments. Environmental Pollution 130, 287-299.

Burkholder, J. M. (1996) Interactions of benthic algae with their substrata. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 253-297). Elsevier, San Diego.

Carpenter, K. D. and Waite, I. R. (2000) Relations

of habitat-specific algal assemblages to land use and water chemistry in Willamette Basin, Oregon. Environmental Monitoring and Assessment 64, 247-257.

Changnon, S. A. and Demissie, M. (1995) Detection of changes in streamflow and floods resulting from climate fluctuations and land use-drainage changes. Climatic Change 32, 411-421.

Chapin, F. S. III, Walker, B. H., Hobbs, R. J., Hooper, D. U., Lawton, J. H., Sala, O. E. and Tilman, D. (1997) Biotic control over the functioning of ecosystems. Science 277, 500-504.

Charles, D. F., Acker, F. W., Hart, D. D., Reimer, C. W. and Cotter, P. B. (2006) Large-scale regional variation in diatom-water chemistry relationships: Rivers of the eastern United States. Hydrobiologia 561, 27-57.

Chen, X., Zhou, W., Pickett, S. T. A., Li, W., Han, L. and Ren, Y. (2016) Diatoms are better indicators of urban stream conditions: A case study in Beijing, China. Ecological Indicators 60, 265-274.

Connell, J.H. (1978) Diversity in tropical rain forests and coral reefs. Science 199, 1302–1310.

Cox, C. B., Moore, P. D. and Ladle, R. J. (2016) Biogeography: An Ecological and Evolutionary Approach. John Wiley & Sons, Ltd, Chichester.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A. and Hess, K. T. (2007) Random forests for classification in ecology. Ecology 88, 2783-2792.

Dar, P. A. and Reshi, Z. A. (2014) Components, processes and concequences of biotic homogenization: A review. Contemporary Problems of Ecology 7, 123-136.

Davis, M. B. and Shaw, R. G. (2001) Range shifts and adaptive responses to quaternary climate change. Science 292, 673-679.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. Ecology 88, 243-251.

Donohue, I., Jackson., A. I., Pusch, M. T. and Irvine, K. (2009) Nutrient enrichment homogenizes lake benthic assemblages at local and regional scales. Ecology 90, 3470-3477.

Elith J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J., Peterson, A., Phillips, S., Richardson, K., Scachetti-Pereira, R., Schapire, R., Soberon, J., Williams, S., Wisz, M. and Zimmermann, N. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Elith J., Leathwick, J. R. and Hastie, T. (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77, 802-813.

Eloranta, P. (1995) Type and quality of river waters in central Finland described using diatom indices. Proceedings of the 13th International Diatom Symposium (ed. by D. Marino and M. Montresor), pp. 271-280. Biopress, Bristol, UK.

Eshleman, K. N. and Hemond, H. F. (1985) The role of organic acids in the adic-base status of surface waters at Bickford Watershed, Massachusetts. Water Resources Research 21, 1503-1510.

Esposito, R. M. M., Horn, S. L., McKnight, D. M., Cox, M. J., Grant, M. C., Spaulding, S. A., Doran, P. T. and Cozzetto, K. D. (2006) Antarctic climate cooling and response of diatoms in glacial meltwater streams. Geophysical Research Letters 33, doi:10.1029/2006GL025903.

Evans, C. D., Monteith, D. T. and Cooper, D. M. (2005) Long-term increases in surface water dissolved organic carbon: Observations, possible causes and environmental impacts. Environmental Pollution 137, 55-71.

Fausch, K. D., Torgersen, C. E., Baxter, C. V. and Li, H. W. (2002) Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. BioScience 52, 483-498.

Fielding, A. H. and Bell, J. F. (1997) A review methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24, 38-49.

Finlay, B. (2002) Global dispersal of free-living microbial eukaryote species. Science 296, 1061-1063.

Finlay, B. J. and Fenchel, T. (2004) Cosmopolitan metapopulations of free-living microbial eukaryotes. Protist 155, 237-244.

Finnish Environment Institute (2013) CORINE Land Cover 20 m. (Available at: https://avaa.tdata.fi/web/paituli).

Foley, J. A., DeFries, R., Asner, G. P., Barford, C., Bonan, G., Carpenter, S. R., Chapin, F. S., Coe, M. T., Daily, G. C., Gibbs, H. K., Helkowski, J. H., Holloway, T., Howard, E. A., Kucharik, C. J., Monfreda, C., Patz, J. A., Prentice, I. C., Ramankutty, N. and Snyder, P. K. (2005) Global consequences of land use. Science 309, 570-574.

Fore, L. S. and Grafe, C. (2002) Using diatoms to assess the biological conditions of large rivers in Idaho (U.S.A). Freshwater Biology 47, 2015-2037.

Friedman J., Hastie T. and Tibshirani, R. (2000) Additive logistic regression: A statistical view of boosting. Annals of Statistics 28, 337-407.

Friedman, J. H. (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29, 1189-1232.

Frissell, C. A., Liss, W. J., Warren, C. E. and Hurley, M. D. (1986) A hierarchical framework for stream habitat classification: viewing streams in a watershed context. Environmental Management 10, 199-214.

Gudmundsdottir, R., Olafsson, J. S., Palsson, S., Gislason, G. M. and Moss, B. (2011) How will increased temperature and nutrient enrichment affect primary producers in sub-Arctic streams? Freshwater Biology 56, 2045-2058.

Guerold, F., Boudot, J.-P., Jacquemin, G., Vein, D.,

Merlet, D. and Rouiller, J. (2000) Macroinvertebrate community loss as a result of headwater stream acidification in the Vosges Mountains (N-E France). Biodiversity & Conservation 9, 767-783.

Hall, R. J., Likens, G. E., Fiance, S. B. and Hendrey, G. R. (1980) Experimental acidification of a stream in the Hubbard Brook experimental forest, New Hampshire. Ecology 61, 976-989.

Hastie, T. and Tibshirani, R. J. (1990) Generalized Additive Models. Chapman & Hall, London.

Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J.-F., Kaufman, D. M., Kerr, J. T., Mittelbach, G. G., Oberdorff, T., O`Brien, E. M., Porter, E. E. and Turner, J. R. G. (2003) Energy, water, and broad-scale geographic patterns of species richness. Ecology 84, 3105-3117.

Heino, J., Virkkala, R. and Toivonen, H. (2009) Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. Biological Reviews 84, 39-54.

Heino, J., Bini, L. M., Karjalainen, S. M., Mykrä, H., Soininen, J., Vieira, L. C. G. and Diniz-Filho, J. A. F. (2010) Geographical patterns of microorganismal community structure: are diatoms ubiquitously distributed across boreal streams? Oikos 119, 129-137.

Heino, J., Bini, L. M., Andersson, J., Bergsten, J., Bjelke, U. and Johansson, F. (2017) Unravelling the correlated of species richness and ecological uniqueness in a metacommunity of urban pond insects. Ecological Indicators 73, 422-431.

Hieber, M., Robinson, C. T., Rushforth, S. R. and Uehlinger, U. (2001) Algal communities associated with different alpine stream types. Arctic, Antarctic, and Alpine Research 33, 447-456.

Hillebrand, H. (2004) On the generality of the latitudinal diversity gradient. The American Naturalist 163, 192-211.

Hillebrand, H. and Blenckner, T. (2002) Regional and local impact on species diversity – from pattern to processes. Oecologia 132, 479-491.

Hoaglund, K. D., Carder, J. P. and Spawn, R. L. (1996) Effects of organic toxic substances. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 469-496). Elsevier, San Diego.

Hough-Snee, N., Kasprak, A., Roper, B. B. and Meredith, C. S. (2014) Direct and indirect drivers of instream wood in the interior Pacific Northwest, USA: decoupling climate, vegetation, disturbance, and geomorphic setting. Riparian Ecology and Conservation 2, 14-34.

Hughes, L. (2000) Biological consequences of global warming: is the signal already apparent? Trends in Ecology & Evolution 15, 56-61.

Hynes, H. (1975) The stream and its valley. Verhandlungen der Internationalen Vereinigung fur theoretische und angewandte Limnologie? 19, 1-15.

IPCC (2014) Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team, R.K. Pachauri and L.A. Meyer (Eds.). IPCC, Geneva, Switzerland.

Jeppesen, E., Kronvang, B., Meerhoff, M., Søndergaard, M., Hansen, K. M., Andersen, H. E., Lauridsen, T. L., Liboriussen, L., Beklioglu, M., Özen, A. and Olesen, J. E. (2009) Climate change effects on runoff, catchment phosphorus loading and lake ecological state, and potential adaptations. Journal of Environmental Quality 38, 1930-1941.

Jing, X., Sanders, N. J., Shi, Y., Chu, H., Classen, A. T., Zhao, K., Chen, L., Shi, Y., Jiang, Y. and He, J.-S. (2015) The links between ecosystem multifunctionality and above- and belowground biodiversity are mediated by climate. Nature Communications 6, 8159.

Johnson, R. K. and Angeler, D. G. (2014) Effects of agricultural land use to stream assemblages: Taxon-specific responses of alpha and beta diversity. Ecological Indicators 45, 386-393.

Juggins S. (2013) Rioja: Analysis of Quaternary Science Data. (Available at: http://cran.r-project.org/web/packages/rioja/index.html).

Jüttner, I., Chimonides, P. D. J., Ormerod, S. J. and Cox, E. J. (2010) Ecology and biogeography of Himalayan diatoms: distribution along gradients of altitude, stream habitat and water chemistry. Fundamental and Applied Limnology 177, 293-311.

Jüttner, I., Reichardt, E. and Ormerod, S. J. (1996) Diatoms as indicators of river quality in the Nepalese Middle Hills with consideration of the effects of habitat-specific sampling. Freshwater Biology 36, 475-486.

Jyrkänkallio-Mikkola, J., Meier, S., Heino, J., Laamanen, T., Pajunen, V., Tolonen, K.T., Tolkkinen, M. and Soininen, J. (2017) Disentangling multiscale environmental effects on stream microbial communities. Journal of Biogeography 44, 1512-1523.

Kolpin, D. W., Furlong, E. T., Mayer, M. T., Thurman, E. M., Zaugg, S. D., Barber, L. B. and Buxton, H. T. (2002) Pharmaceuticals, hormones, and other organic wastewater contaminants in U.S. streams, 1999-2000: a national reconnaissance. Environmental Science & Technology 36, 1202-1211.

Krammer, K. and Lange-Bertalot, H. (1986–1991) Bacillariophyceae. Süβwasserflora von Mitteleuropa 2 (1-4). Gustav Fischer Verlag, Stuttgart.

Kryza, M., Werner, M., Dore, A. J., Blas, M. and Sobik, M. (2012) The role of annual circulation and precipitation on national scale deposition of atmospheric sulphur and nitrogen compounds. Journal of Environmental Management 109, 70-79.

Lake, P. S. (2000) Disturbance, patchiness, and diversity in streams. Journal of North American Benthological Society 19, 573-592.

Lange-Bertalot, H. and Metzeltin, D. (1996) Iconographica diatomologica, Volume 2. Indicators of oligotrophy. 800 taxa representative of three ecologically distinct lake types: carbonate buffered, oligodystrophic, weakly buffered soft water. Koeltz Scientific Books, Koenigstein.

Larras, F., Coulaaud, R., Gautreau, E., Billoir, E., Rosebery, J. and Usseglio-Polatera, P. (2017) Assessing anthropogenic pressures on streams: a random forest approach based on benthic diatom communities. Science of the Total Environment 15, 1101-1112.

Lavoie, I., Campeau, S., Grenier, M. and Dillon, P. J. (2006) A diatom-based index for the biological assessment of eastern Canadian rivers: an application of correspondence analysis (CA). Canadian Journal of Fisheries and Aquatic Sciences 63, 1793-1811.

Lefcheck, J. S. (2016) PiecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. Methods in Ecology and Evolution 7, 573-579.

Legendre, P. and De Cáceres, M. (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecology Letters 16, 951-963.

Leland, H. V. and Porter, S. D. (2000) Distribution of benthic algae in the upper Illinois River basin in relation to geology and land use. Freshwater Biology 44, 279-301.

Levesque, D., Hudon, C., James, P. M. A. and Legendre, P. (2017) Environmental factors structuring benthic primary producers at different spatial scales in the St. Lawrence River (Canada). Aquatic Sciences 79, 345-356.

Lewin, J. C. and Lewin, R. A. (1960) Auxotrophy and heterotrophy in marine littoral diatoms. Canadian Journal of Microbiology 6, 127-134.

Liess, A., Lange, K., Schulz, F., Piggott, J. J., Matthaei, C. D. and Townsend, C. R. (2009) Light, nutrients and grazing interact to determine diatoms species richness via changes to productivity, nutrient state and grazing activity. Journal of Ecology 97, 326-336.

Liu, S., Xie, G., Wang, L., Cottenie, K., Liu, D. and Wang, B. (2016) Different roles of environmental variables and spatial factors in structuring stream benthic diatom and macroinvertebrate in Yangtze River Delta, China. Ecological Indicators 61, 602-611.

Lobo, E. A., Katoh, K. and Aruga, Y. (1995) Response of epilithic diatom assemblages to water pollution in rivers in the Tokyo Metropolitan area, Japan. Freshwater Biology 34, 191-204.

Lowe, R. L. and Pan, Y. (1996) Benthic algal communities as biological monitors. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 705-739). Elsevier, San Diego.

Maloney, K. O. and Weller, D. E. (2011) Anthropogenic disturbance and streams: land use and land-use change affect stream ecosystems via multiple pathways. Freshwater Biology 56, 611-626.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., Zingone, A. and Bowler, C. (2016) Insights into global diatom distribution and diversity in the world's ocean. Proceedings of the National Academy of Sciences of the United States of America 113, E1516-E1525.

Martiny, J. B. H., Bohannan, J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P.J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H. and Staley, J. T. (2006) Microbial biogeography: putting microorganisms on the map. Nature Reviews Microbiology 4, 102-112.

Martiny, J. B. H., Eisen, J. A., Penn, K., Allison, S. D. and Horner-Devine, M. C. (2011) Drivers of bacterial beta-diversity depend on spatial scale. Proceedings of the National Academy of Sciences of the United States of America 108, 7850-7854.

McCarty, J. P. (2001) Ecological consequences of recent climate change. Conservation Biology 15, 320-331.

McCormick, P. V. (1996) Resource competition and species coexistence in freshwater benthic algal assemblages. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 229-252). Elsevier, San Diego.

McCullagh, P. and Nelder, J. A. (1989) Generalized Linear Models. Chapman & Hall, Florida.

Michels, A., Umana, G. and Raeder, U. (2006) Epilithic diatom assemblages in rivers draining into Golfo Dulce (Costa Rica) and their relationship to water chemistry, habitat characteristics and land use. Archiv fur Hydrobiologie 165, 167-190.

Mittelbach, G. G. (2012) Community Ecology. Sinauer Associates, Inc., Sunderland.

Moravcova, A., Rauch, O., Lukavsky, J. and Nedbalova, L. (2013) The responses of epilithic diatom assemblages to sewage pollution in mountain streams of the Czech Republic. Plant Ecology and Evolution 146, 153-166.

Moss, B. (1998) Ecology of Fresh Waters. Blackwell Science, Oxford.

National Land Survey of Finland (2013) Elevation model 10 m. (Available at: https://avaa.tdata.fi/web/paituli).

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Billinski, T. M., Stanish, L. F., Knelman, J. E., Darcy, J. L., Lynch, R. C., Wickey, P. and Ferrenberg, S. (2013) Patterns and processes of microbial community assembly. Microbiology and Molecular Biology Reviews 77, 342-356.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre,

P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. and Wagner, H. (2015) Vegan: Community Ecology Package. (Available at: http://cran.r-project.org/web/packages/vegan/index.html).

Oliveira, L. and Huynh, H. (1990) Phototrophic growth of microalgae with allantoic acid or hypoxanthine serving as nitrogen source, implications for purine-N utilization. Canadian Journal of Fisheries and Aquatic Sciences 47, 351-356.

Overpeck J. T., Webb III T. and Prentice I. C. (1985) Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs. Quaternary Research 23, 87-108.

Palmer, M. A. and Poff, N. L. (1997) The influence of environmental heterogeneity on patterns and processes in streams. Journal of the North American Benthological Society 16, 169-173.

Parmesan, C. (2006) Ecological and evolutionary responses to recent climate change. Annual Review of Ecology, Evolution, and Systematics 37, 637-669.

Parmesan, C. and Yohe, G. (2003) A globally coherent fingerprint of climate change impact across natural systems. Nature 421, 37-42.

Passy, S. I. (2001) Spatial paradigms of lotic diatom distribution: a landscape ecology perspective. Journal of Phycology 37, 370-378.

Passy, S. I. (2008) Continental diatom biodiversity in stream benthos declines as more nutrients become limiting. Proceedings of the National Academy of Sciences of the United States of America 105, 9663-9667.

Passy, S. I. (2009) The relationship between local and regional diatom richness is mediated by the local and regional environment. Global Ecology and Biogeography 18, 383-391.

Passy. S. I. (2010) A distinct latitudinal gradient of diatom diversity is linked to resource supply. Ecology 91, 36-41.

Patrick, R. (1971) The effects of increasing light and temperature on the structure of diatom communities. Limnology and Oceanology 16, 405-421.

Peterson, C. G. (1996) Response of benthic algal communities to natural physical disturbance. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 375-402). Elsevier, San Diego.

Piggott, J. J., Salis, R. K., Lear, G., Townsend, C. R. and Matthaei, C. D. (2015) Climate warming and agricultural stressors interact to determine stream periphyton community composition. Global Change Biology 21, 206-222.

Planas, D. (1996) Acidification effects. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 497-530). Elsevier, San Diego.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E. and Stromberg, J. C. (1997) The natural flow regime: a paradigm for river conservation and restoration. BioScience 47, 769-784.

Potapova, M. G. and Charles, D. F. (2002) Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. Journal of Biogeography 29, 167-187.

Potapova, M. and Charles, D. F. (2003) Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. Freshwater Biology 48, 1311-1328.

Potapova M. and Charles D. F. (2007) Diatom metrics for monitoring eutrophication in rivers of the United States. Ecological Indicators 7, 48-70.

Potapova, M. G., Charles, D. F., Ponader, K. C. and Winter, D. M. (2004) Quantifying species indicator values for trophic diatom indices: a comparison of approaches. Hydrobiologia 517, 25-41.

Pound, K. L., Lawrence, G. B. and Passy, S. I. (2013) Wetlands serve as natural sources for improvement of stream ecosystem health in regions affected by acid deposition. Global Change Biology 19, 2720-2728.

Prygiel, J., Whitton, B. A. and Bukowska, J. (1997) Use of algae for monitoring rivers III. Agence de l`Eau Artois-Picardie, Douai.

R Development Core Team. (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (Available at: https://www.r-project.org/).

Rahel, F. J. (2002) Homogenization of freshwater faunas. Annual Review of Ecology and Systematics 33, 291-315.

Rai, L. C., Gaur, J. P. and Kumar, H. D. (1981) Phycology and heavy-metal pollution. Biological Reviews 56, 99-151.

Ridgeway G. (2010) gbm: Generalized Boosted Regression Models. (Available at: http://cran.r-project.org/web/packages/gbm/index.html).

Rimet F., Cauchie H., Hoffmann L. and Ector L. (2005) Response of diatom indices to simulated water quality improvements in a river. Journal of Applied Phycology 17, 119-128.

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A. and Bouchez, A. (2016) R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database: The Journal of Biological Databases and Curation, DOI: 10.1093/database/baw016.

Rose D. T. and Cox E. J. (2014) What constitutes Gomphonema parvulum? Long-term culture studies show that some varieties of G. parvulum belong with other Gomphonema species. Plant Ecology and Evolution 147, 366–373.

Rothwell, J. J., Dise, N. B., Taylor, K. G., Allott, T. E. H., Scholefield, P., Davies, H. and Neal, C. (2010) A spatial and seasonal assessment of river water chemistry across North West England. Science of

the Total Environment 408, 841-855.
Rott E., Cantonati M., Füreder L. and Pfister P. (2006) Benthic algae in high altitude streams of the Alps - a neglected component of the aquatic biota. Hydrobiologia 562, 195-216.
Round, F. E. (2004) pH scaling and diatom distribution. Diatom 20, 9-12.
Round, F. E., Crawford, R. M., and Mann, D. G. (1990) The Diatoms: Biology and Morphology of the Genera. Cambridge University Press, Cambridge.
Sandin, L. and Verdonschot, P. F. M. (2006) Stream and river typologies – major results and conclusions from the STAR project. Hydrobiologia 566, 33-37.
Schneck, F., Lange, K., Melo, A. S., Townsend, C. R. and Matthaei, C. D. (2017) Effects of a natural flood disturbance on species richness and beta diversity of stream benthic diatom communities. Aquatic Ecology, 1-13.
Simpson G. L. and Oksanen J. (2013) Analogue: Analogue and Weighted Averaging Methods for Palaeoecology. (Available at: http://cran-r-project.org/web/packages/analogue/index.html).
Smol J. P. (2010) The power of the past: using sediments to track the effects of multiple stressors on lake ecosystems. Freshwater Biology 55 (Suppl. 1), 43-59.
Smol, J. P. and Stoermer, E. F. (2010) The Diatoms: Applications for the Environmental and Earth Sciences. Cambridge University Press, New York.
Smucker, N. J. and Vis, M. L. (2013) Can pollution severity affect diatom succession in streams and could it matter for stream assessments? Journal of Freshwater Ecology 28, 329-338.
Snelder, T. H. and Biggs, B. J. F. (2002) Multiscale river environment classification for water resources management. Journal of the American Water Resources Association 38, 1225-1239.
Soininen, J. (2004) Determinants of benthic diatom community structure in boreal streams: the role of environmental and spatial factors at different scales. International Review of Hydrobiology 89, 139-150.
Soininen J. (2007) Environmental and spatial control of freshwater diatoms - a review. Diatom Research 22, 473-490.
Soininen, J., Jamoneau, A., Rosebery, J. and Passy, S. I. (2016) Global patterns and drivers of species and trait composition of diatoms. Global Ecology and Biogeography 25, 940-950.
Soininen, J., Korhonen, J. J. and Luoto, M. (2013) Stochastic species distributions are driven by organism size. Ecology 94, 660-670.
Soininen J. and Niemelä P. (2002) Inferring the phosphorus levels of rivers from benthic diatoms using weighted averaging. Archiv Fur Hydrobiologie 154, 1-18.
Soininen, J., Paavola, R. and Muotka, T. (2004) Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. Ecography 27, 330-342.
Steinman, A. D. (1996) Effects of grazers on freshwater benthic algae. In R. J. Stevenson, M. L. Bothwell, M. L. & R. L. Lowe (Eds.), Algal Ecology: Freshwater Benthic Ecosystems (pp. 341-373). Elsevier, San Diego.
Stenger-Kovacs, C., Toth, L., Toth, F., Hajnal, E. and Padisak, J. (2014) Stream order-dependent diversity metrics of epilithic diatom assemblages. Hydrobiologia 721, 67-75.
Stevenson, R. J. (1997) Scale-dependent determinants and consequences of benthic algal heterogeneity. Journal of North American Benthological Society 16, 248-262.
Stevenson, R. J., Bothwell, M. L. and Lowe, R. L. (1996) Algal Ecology: Freshwater Benthic Ecosystems. Elsevier, San Diego.
Stoodley, P., Sauer, K., Davies, D. G. and Costerton, J. W. (2002) Biofilms as complex differentiated communities. Annual Review of Microbiology 56, 187-209.
Sutherland, A. B., Meyer, J. L. and Gardiner, E. P. (2002) Effects of land cover on sediment regime and fish assemblage structure in four southern Appalachian streams. Freshwater Biology 47, 1791-1805.
Taka, M., Kokkonen, T., Kuoppamäki, K., Niemi, T., Sillanpää, N., Valtanen, M., Warsta, L. and Setälä, H. (2017) Spatio-temporal patterns of major ions in urban stormwater under cold climate. Hydrological Processes 31, 1564-1577.
Tang, T., Wu, N., Li, F., Fu, X. and Cai, Q. (2013) Disentangling the roles of spatial and environmental variables in shaping benthic algal assemblages in rivers of central and northern China. Aquatic Ecology 47, 453-466.
Teittinen, A., Taka, M., Ruth, O. and Soininen, J. (2015) Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. Science of the Total Environment 530, 279-289.
ter Braak, C. J. F. and Juggings, S. (1993) Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia, 269/270, 485-502.
ter Braak J. F. and van Dam H. (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178, 209-223.
Thuiller, W., Georges, D., Engler, R. and Breiner, F. (2016) biomod2: Ensemble Platform for Species Distribution Modeling. (Available at: mran.microsoft.com/package/biomod2/biomod2.pdf).
Thuiller, W., Lafourcade, B., Engler, R. and Araújo, M. B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32, 369-373.
Tilman, D. (1977) Resource competition between

plankton algae: an experimental and theoretical approach. Ecology 58, 338-348.

Tudesque, L., Tisseuil, C. and Lek, S. (2014) Scale-dependent effects of land cover on water physicochemistry and diatom-based metrics in a major river system, the Adour-Garonne basin (South Westerns France). Science of the Total Environment 466, 47-55.

Urrea G. and Sabater S. (2009) Epilithic diatom assemblages and their relationship to environmental characteristics in an agricultural watershed (Guadiana River, SW Spain). Ecological Indicators 9, 693-703.

van Dam, H., Mertens, A. and Sinkeldam, J. (1994) A coded checklist and ecological indicators values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology 28, 117-133.

Van de Vijver, B. and Beyens, L. (1999) Biogeography and ecology of freshwater diatoms in Subantarctica: a review. Journal of Biogeography 26, 993-1000.

Vannote, R. L. M., Minshall, G. W., Cummins, K. W., Sedell, J. R. and Cushing, C. E. (1980) The river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences 37, 130-137.

Vanormelingen, P., Verleyen, E. and Vyverman, W. (2008) The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. Biodiversity and Conservation 17, 393-405.

Varanka, S. and Luoto, M. (2012) Environmental determinants of water quality in boreal rivers based on partitioning methods. River Research and Applications 28, 1034-1046.

Venäläinen, A. and Heikinheimo, M. (2002) Meteorological data for agricultural applications. Physics and Chemistry of the Earth 27, 1045-1050.

Verleyen, E., Vyverman, W., Sterken, M., Hodgson, D. A., De Wever, A., Juggins, S., Van de Vijver, B., Jones, V. J., Vanormelingen, P., Roberts, D., Flower, R., Kilroy, C., Souffreau, C. and Sabbe, K. (2009) The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. Oikos 118, 1239-1249.

Vilmi, A., Karjalainen, S. M. and Heino, J. (2017) Ecological uniqueness of stream and lake diatom communities shows different macroecological patterns. Diversity and Distributions 23, 1042-1053.

Virtanen, L. and Soininen, J. (2012) The roles of environment and space in shaping stream diatom communities. European Journal of Phycology 47, 160-168.

Vyverman, W., Verleyen, E., Sabbe, K., Vanhoutte, K., Sterken, M., Hodgson, D. A., Mann, D. G., Juggins, S., Van de Vijver, B., Jones, V., Flower, R., Roberts, D., Chepurnov, V. A., Kilroy, C., Vanormelingen, P. and De Wever, A. (2007) Historical processes constrain patterns in global diatom diversitymen. Ecology 88, 1924-1931.

Walker, C. E. and Pan, Y. D. (2006) Using diatom assemblages to assess urban stream conditions. Hydrobiologia 561, 179-189.

Walsh, G. and Wepener, V. (2009) The influence of land use on water quality and diatom community structures in urban and agriculturally stressed rivers. Water SA 35, 579-594.

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., Fromentin, J.-M., Hoegh-Guldberg, O. and Bairlein, F. (2002) Ecological responses to recent climate change. Nature 416, 389-395.

Wang, J., Meier, S., Soininen, J., Casamayor, E. O., Pan, F., Tang, X., Yang, X., Zhang, Y., Wu, Q., Zhou, J. and Shen, J. (2017) Regional and global elevational patterns of microbial species richness and evenness. Ecography 40, 393-402.

Wang, L., Brenden, T., Seelbach, P., Cooper, A., Allan, D., Clark, R. Jr. and Wiley, M. (2008) Landscape based identification of human disturbance gradients and reference conditions for Michigan streams. Environmental Monitoring and Assessment 141, 1-17.

Weckström, J., Korhola, A. and Blom, T. (1997a) The relationship between diatoms and water temperature in thirty Fennoscandian lakes. Arctic and Alpine Research 29, 75-92.

Weckström, J., Korhola, A. and Blom, T. (1997b) Diatoms as quantitative indicators of pH and water temperature in subarctic Fennoscandian lakes. Hydrobiologia 347, 171-184.

Weilhoefer, C. L. and Pan, Y. D. (2006) Diatom assemblages and their associations with environmental variables in Oregon Coast Range streams, USA. Hydrobiologia 561, 207-219.

Whitton, B.A., Rott, E. and Friedrich, G. (1991) Use of algae for monitoring rivers. Institut für Botanik, Universität Innsbruck, Innsbruck.

Yu, S. F. and Lin, H. J. (2009) Effects of agriculture on the abundance and community structure of epilithic algae in mountain streams of subtropical Taiwan. Botanical Studies 50, 73-87.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. and Gemeinholzer, B. (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources 15, 526-542.

Zobel, M. (1997) The relative role of species pools in determining plant species richness: an alternative explanation of species coexistence? Trends in Ecology and Evolution 12, 266-269.

Paper I

Pajunen, V., Luoto, M., Soininen, J. 2016. Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography* 25, 198-206.

Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2016) **25**, 198–206

Global Ecology and Biogeography
A Journal of Macroecology

RESEARCH PAPERS

Climate is an important driver for stream diatom distributions

Virpi Pajunen *, Miska Luoto and Janne Soininen

Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland

*Correspondence: Virpi Pajunen, Department of Geosciences and Geography, University of Helsinki, PO Box 64, FI-00014 University of Helsinki, Helsinki 00014, Finland.
E-mail: virpi.pajunen@helsinki.fi

ABSTRACT

Aim The species distributions of macroorganisms are widely studied, yet microbial distributions at species level remain poorly resolved. We explored the relative contributions of climatic and local environmental factors in explaining the distribution patterns of unicellular diatoms.

Location A geographical gradient of *c.* 1200 km in Finland (60°–70° N).

Methods We modelled the distributions of 157 diatom taxa sampled in 227 stream sites using climatic and local environmental predictors with four different modelling techniques: generalized linear models, generalized additive models, boosted regression trees and Random Forest. We used models with three separate sets of predictors: environment only, climate only and the full set of predictors. The model performances were evaluated using the area under the curve of a receiver operating characteristic plot and the true skill statistic values by a four-fold cross-validation approach.

Results We found that the predictive performance of the full models was highest, indicating the importance of both the local environment and large-scale climatic factors in diatom distribution patterns. However, climate-only models outcompeted the environment-only models in predicting diatom distributions. The explanatory variables had varying importance across species and growing degree days and precipitation had the highest relative importance in the full models. We also found that the predictability of the distributions varied greatly among species, but the differences among families were typically small.

Main conclusions Our results suggest that at a broad geographical scale climate-related factors are important determinants of diatom distributions and may be stronger drivers than local variables. The inclusion of both climatic and local environmental factors in species distribution modelling facilitates the understanding of the joint effects of these drivers on microorganisms in future conditions. From an applied perspective, our study demonstrated that species distribution models serve as an important tool in explaining and predicting microbial distributions.

Keywords
Climate, diatoms, microorganisms, species distribution models, streams.

INTRODUCTION

Climate is known to have a strong influence on species distributions on Earth (Davis & Shaw, 2001; Walther *et al.*, 2002). The climatic effect typically exceeds the importance of land-cover variables, for example, in determining species distributions at both continental and regional scales (Pearson *et al.*, 2004; Venier *et al.*, 2004; Luoto *et al.*, 2007). Species distribution models (SDMs) statistically relate the geographical distribution of species to their present environment and are useful tools for exploring both the current and future distribution of species (Guisan & Zimmermann, 2000). More realistic and complex models are being continuously developed for the use in biogeography, conservation biology, climate change research and

DOI: 10.1111/geb.12399

species or habitat management (Guisan & Thuiller, 2005; Thuiller *et al.*, 2008). Although SDMs have been widely used for macroorganisms, comparable distribution models for microorganisms are still rare (Ladau *et al.*, 2013; Soininen & Luoto, 2014).

The traditional view suggests that microbial taxa are ubiquitous (Finlay, 2002) and lack any biogeographical patterns associated with evolution, dispersal limitation or glaciation history. Thus, their distributions would be governed by local environmental variables that may be spatially autocorrelated. However, there is growing evidence that microbial taxa also have biogeographical patterns largely similar to those documented for macroorganisms (Martiny *et al.*, 2006; Astorga *et al.*, 2012; Nemergut *et al.*, 2013). Despite this, the use of SDMs for microbial distributions is still rare (but see Ladau *et al.*, 2013; Soininen *et al.*, 2013). One of the reasons why SDMs are not often employed for microorganisms is simply the scarcity of large-scale microbial datasets due to the fact that microbe biogeography has only recently become a focus of attention (Martiny *et al.*, 2006).

The predictability of microbial distributions probably differs widely between broad taxonomic groups because different microbial taxa may differ in their dispersal capabilities, extinction–colonization dynamics and range sizes, and environmental tolerance can also vary between species within the same taxonomic group (Vanormelingen *et al.*, 2008). Diatoms (Bacillariophyta), a large and diverse group of algae, inhabit all types of aquatic ecosystems globally. The conventional view is that diatoms respond merely to local environmental variables such as water chemistry and physical variables; they are therefore widely used as bioindicators (Battarbee, 2000; Soininen *et al.*, 2004). However, diatom richness and composition may also be related to habitat availability and geographical factors (Soininen *et al.*, 2004; Telford *et al.*, 2006). Recently, large-scale historical factors (i.e. glaciation history, speciation) have been suggested to explain more of the global geographical distributions of diatom genera than current environmental conditions (Vyverman *et al.*, 2007). To understand diatom distributions it is essential to recognize whether distributions are constrained by climatic variables and limited dispersal or only by local environmental conditions (Potapova & Charles, 2002; Bennett *et al.*, 2010).

Here, we use SDMs to examine whether diatom distributions are governed not only by local environmental variables but also by large-scale climatic factors. We use an extensive dataset of 227 sites encompassing the latitudinal range of 1200 km in Finland and model the distributions of 157 benthic diatom taxa using local environmental and climatic variables as predictors. The aim of this study is three-fold. We first examine if the predictive performance of climate-only models is higher than that of environment-only models. Second, we investigate if the predictive performance of climate–environment models (i.e. full models) is higher than environment-only and climate-only models. Third, we examine the variable importance of local environmental and climatic variables for the distribution of diatoms. These analyses would indicate whether climatic variables are also needed for the reliable modelling of diatom species distributions and add more insights into the key drivers of microbial distributions in general.

MATERIALS AND METHODS

Data collection

The present data set was obtained by combining three diatom data sets collected in Finland between 1986 and 2004 (227 sites in total) (Fig. 1). Although sampling occasions cover a wide temporal range, we consider these samples to be comparable because the sampling methods were identical. Furthermore, Korhonen *et al.* (2013) have demonstrated that although diatom assemblages vary in time, compositional differences between assemblages growing in different types of environments, for example oligotrophic versus eutrophic waters, remain significant.

The first data set comprised 56 sites sampled by Eloranta (1995) in 1986. These sites mainly represent near-pristine conditions, being only marginally affected by agriculture, forestry and fish farming and they are located in central Finland. The second data set covered 141 sites that were distributed from southern to northern Finland including nearly pristine and human-impacted sites (Soininen *et al.*, 2004). The sampling was conducted between 1996 and 2001. Finally, two data sets were combined with a set of 30 pristine sites sampled in July 2004 in northern Finnish Lapland. All sampling was performed during low-flow conditions in July and August. The whole data set covered long gradients in conductivity, pH, humus and nutrient concentrations (see Appendix S1 in Supporting Information).

At each sampling site, the minimum of five replicate pebble-to-cobble (5–15 cm) sized stones was collected. Diatoms were sampled by brushing stones with a toothbrush, according to the recommendations of Kelly *et al.* (1998). At most of the sites, water samples were taken simultaneously with diatom samples. They were analysed for total phosphorus (TP), pH, conductivity and water colour using national standards. For some of the sites (< 20%), water chemistry data were taken from the national water quality database, using results from the nearest sampling occasion. Current velocity, shading by the canopy and stream width were measured at each site along five transects per site perpendicular to the flow and covering the whole study section. Diatom samples were cleaned from organic material in the laboratory using wet combustion with acid (HNO_3:H_2SO_4; 2:1) and mounted in Naphrax or Dirax. A total of 250–500 frustules per sample were identified to species level according to Krammer & Lange-Bertalot (1986–1991) and Lange-Bertalot & Metzeltin (1996) and counted using phase contrast light microscopy (magnification 1000×) by two analysts who harmonized species identification.

Climatic variables

Three climatic variables were chosen for modelling the climatic effect on diatom species distribution. These variables were

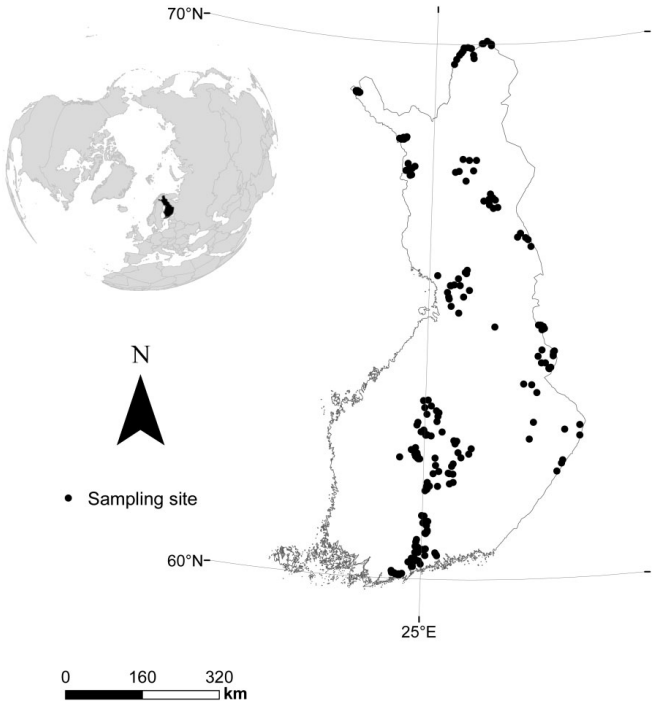


Figure 1 Location of the sampling sites ($n = 227$) in Finland, northern Europe. The index map represents the location of Finland in the Northern Hemisphere.

growing degree days (GDD) (range 273–1308 GDD), growing season precipitation sum from May to September (PRECS) (range 254–347 mm) and water balance (WAB) (range 242–403 mm) (Appendix S1). GDD was chosen as it represents the energy requirements of the species, while PRECS and WAB represent the moisture availability in the environment, which in running waters is connected to the extent of recharge and runoff. The temperature limit in GDD was adjusted to 5 °C. WAB was calculated according to Skov & Svenning (2004) by summing up the monthly differences between precipitation and potential evapotranspiration (PET). Monthly PET was calculated as $PET = 58.93 \times T_{bio}/12$ where T_{bio} is the Holdridge biotemperature, defined as the annual mean of monthly temperatures with negative monthly values adjusted to zero (Holdridge, 1967; Lugo *et al.*, 1999). The climatic data set covered the years 1981–2010 and was obtained from the Finnish Meteorological Institute. Multiple linear regression was used to relate climate data to the latitude, longitude and elevation of each study site, downscaling climate data from a 10 km × 10 km resolution grid to the study site (Finnish Meteorological Institute; Venäläinen & Heikinheimo, 2002). The local and climatic variables were tested for covariance with the nonparametric Spearman's rank correlation coefficient. All predictor variables had relatively low collinearity [$r_s < |0.70|$ (Appendix S2)].

Data analyses and modelling

Diatom species occurring in at least 5% and a maximum of 95% of the sites were included in the statistical analyses. Collectively, the data comprised presence–absence records of 157 taxa from 227 sampling sites. In addition to three climatic variables, three local environmental variables (TP, conductivity and water colour) were used to explain the diatom distributions. These variables were chosen because of their strong impact on diatom species occurrences (Soininen *et al.*, 2004) and they are widely used variables in corresponding studies (e.g. Astorga *et al.*, 2012). We used TP instead of dissolved phosphorus as total nutrient concentrations are better preserved in remote field conditions. Moreover, the concentrations of TP and dissolved phosphorus showed high correlation in the study area ($r_p = 0.75$, $n = 172$ samples). We included conductivity as an alternative to water pH as it is more conservative and has been identified as a stronger explanatory variable for diatom community structure than pH in boreal streams (Soininen *et al.*, 2004). Water colour was used to indicate the amount of dissolved organic carbon in the water because in boreal regions water colour typically originates from humic compounds (Steinberg, 2003).

Three sets of diatom species distribution models were conducted for each species: environment-only, climate-only and the

full model. In the environment-only model, species distribution was modelled only by local environmental variables, whereas in the climate-only model only climatic variables were used. All six variables were included in the full model.

All distribution models were applied via the BIOMOD framework (Thuiller *et al.*, 2009) fitted in R (version 3.1.1; R Development Core Team, 2014). We used four different modelling algorithms to guard against potential differences related to methodologies (Elith *et al.*, 2008): a generalized linear model (GLM), a generalized additive model (GAM), boosted regression trees (BRT) and Random Forest (RF). GLMs are mathematical extensions of linear models which allow nonlinearity and non-constant variance structures in the data, whereas GAMs, nonparametric extensions of GLMs, estimate the form of the relationship between a response variable and predictors using smoothers (Yee & Mitchell, 1991). The machine learning techniques, BRT and RF, are highly efficient at fitting nonparametric data, can manage various types of predictor variables, do not require prior data transformation and automatically take into account interaction effects between predictors (Cutler *et al.*, 2007; Elith *et al.*, 2008). These methods have been shown to have many advantages, for example a smaller prediction error, compared with GLMs and GAMs (Elith *et al.*, 2006; Cutler *et al.*, 2007; De'ath, 2007). The BRT method, in particular, is able to reduce both bias and noise in the data (Elith *et al.*, 2008). The principles of these modelling algorithms have been described in more detail in previous literature: McCullagh & Nelder (1989) (GLMs), Hastie & Tibshirani (1990) (GAMs), Friedman (2001), De'ath (2007) and Elith *et al.* (2008) (BRTs) and Breiman (2001) (RF).

Model performance was assessed using a cross-validation (CV) approach: the models were fitted four times by using a random sample of 70% of the data and subsequently evaluated against the remaining 30%. At each CV run, the predicted and observed occurrences of species were compared by calculating the area under the curve of a receiver operating characteristic plot (AUC) (Fielding & Bell, 1997) and true skill statistics (TSS) (Allouche *et al.*, 2006). The CV approach robustly accounts for possible non-independence (i.e. spatial autocorrelation) of the data (Hijmans, 2012). AUC provides an evaluation of the agreement between the observed presence/absence records over a range of probability thresholds above which the model predicts presence (Fielding & Bell, 1997). TSS considers sensitivity, i.e. the ability to identify taxon presence, and specificity, i.e. the ability to identify absence, and is independent of prevalence (Allouche *et al.*, 2006). The models have at least intermediate predictive performance if AUC values are > 0.7 (following Swets, 1988) and TSS values are > 0.4 (following Landis & Koch, 1977).

The chosen sets of environmental variables were included in the models. In the GLMs the quadratic terms of the predictors were included to examine the probability of curvilinear relationships between the response variables and predictors (Crawley, 2007). GLMs and GAMs were fitted for the binomial distribution of errors and a logit link function was applied. In the GAMs the initial degrees of smoothness were set to 4. BRTs were performed with a maximum number of 3000 trees, an interaction depth of 6 and a learning rate of 0.001. In RFs, the number of trees (k) was set to 500 and the minimum size of terminal nodes was set to 5.

Predicted probabilities of occurrence were converted to presence/absence predictions using the threshold value maximizing sensitivity and specificity (Liu *et al.*, 2005; Levinsky *et al.*, 2013). The difference in model performances between the three sets of SDMs was tested with the nonparametric Wilcoxon signed rank test, where a significant difference between the mean ranks of the compared test pair is indicated by a P-value < 0.05.

For each species, the importance of each variable in the models was assessed in BIOMOD by randomizing each variable individually and then projecting the model with the randomized variable while keeping the other variables unchanged. The predictions of the model containing the randomized variable were then correlated with those of the original models. Finally, the importance of the variable was calculated as one minus the correlation; higher values indicate predictors that are more important for the model (Thuiller *et al.*, 2009). This analysis was repeated 10 times for each modelling technique and the resulting variable importance values were averaged.

The variation in the species occurrence data among the two predictor groups, climatic and local environmental, was decomposed using the variation partitioning approach based on redundancy analysis (RDA) applying the R package VEGAN (Oksanen *et al.*, 2015). The variation was decomposed in two steps: first, variable groups contained the three climatic and three local environmental variables. Second, climatic group was supplemented with mean temperature of the coldest month and local environmental group with shading and current velocity.

RESULTS

All four modelling techniques had similar patterns in predictive performances for SDMs: climate-only models had significantly higher AUC (Wilcoxon test, all $P = 0.000$) and TSS (all $P = 0.000$) values than environment-only models (Fig. 2, Appendix S3). For all modelling techniques, full models had a significantly higher predictive ability than environment-only models. Moreover, the predictive performance of the full BRT and RF models was significantly higher than that of climate-only models.

Overall, the full models had the best predictive performance [AUC medians ranged from 0.777 (GLM) to 0.808 (RF), TSS medians ranged from 0.556 (GLM) to 0.601 (BRT); Appendix S4]. Furthermore, the predictive performances of the models varied widely among species (e.g. in full models for BRT, AUC values ranged from 0.411 to 0.997 depending on the species; Appendix S4). However, no significant differences among diatom families were found with respect to their predictive performances (Kruskal–Wallis test, all $P > 0.01$).

Climatic variables overall had a higher variable importance than local environmental variables in the full models. The only exception was conductivity, which had the second or third

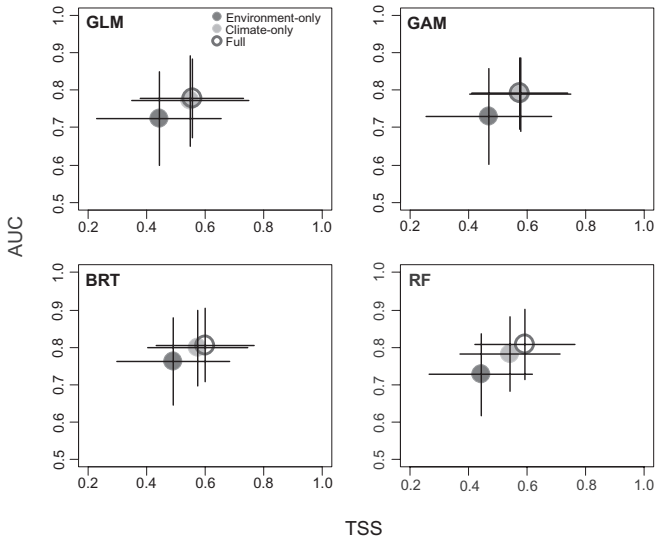


Figure 2 Predictive performances for three sets of diatom species distribution models presented as the area under the curve of a receiver operating characteristic plot (AUC) and the true skill statistic (TSS) values separately for the four modelling techniques used: generalized linear modelling (GLM), generalized additive modelling (GAM), boosted regression trees (BRT) and random forest (RF). The environment-only models consist of three local environmental predictors (total phosphorus, conductivity and water colour), the climate-only models consist of three climatic predictors (growing degree days, summer precipitation sum and water balance) and full models consist of all six predictors. The lines represent standard deviation and the crossing point of the lines represents the median.

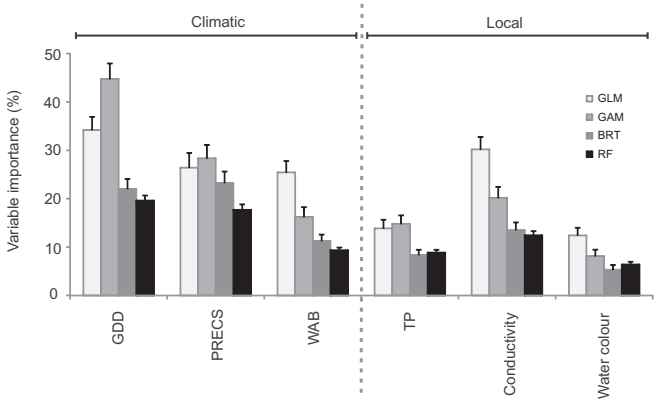


Figure 3 Variable importance (%), i.e. the relative influence, of climatic and local environmental variables for diatom species ($n = 157$) distributions using four different modelling methods: generalized linear modelling (GLM), generalized additive modelling (GAM), boosted regression trees (BRT) and random forest (RF). Error bars represent standard errors. GDD, growing degree days; PRECS, summer precipitation sum; WAB, water balance; TP, total phosphorus.

highest importance depending on the modelling technique (Fig. 3). The variables that had the highest relative importance on distributions were GDD (for GLM, BRT and RF) and PRECS (for GAM), whereas TP and water colour had the lowest relative importance. The variable importance of each predictor varied widely among different species (Fig. 3), but the differences among diatom families were not statistically significant (Kruskal–Wallis test, all $P > 0.01$).

In RDA-based variation partitioning, the three climatic variables used in SDMs (8%) explained more of the total variation than the three local environmental variables (5%) or their joint effects (7%). The majority (80%) of the total variation was left undetermined (Appendix S5). The addition of three new variables (temperature of the coldest month, shading and current velocity) into the variation partitioning did not change the percentages of total variation for climatic or local environmental groups but increased their joint effects to 9%.

As examples of predictive distribution modelling, occurrences of the species *Achnanthes pusilla* (Grunow) were well predicted in the study area (Fig. 4). However, for the species *Eunotia implicata* (Nörpel, Lange-Bertalot & Alles) the environment-only model predicted several false presences, as also shown by the low AUC value (0.681), while the predictive ability was clearly higher for climate-only and full models.

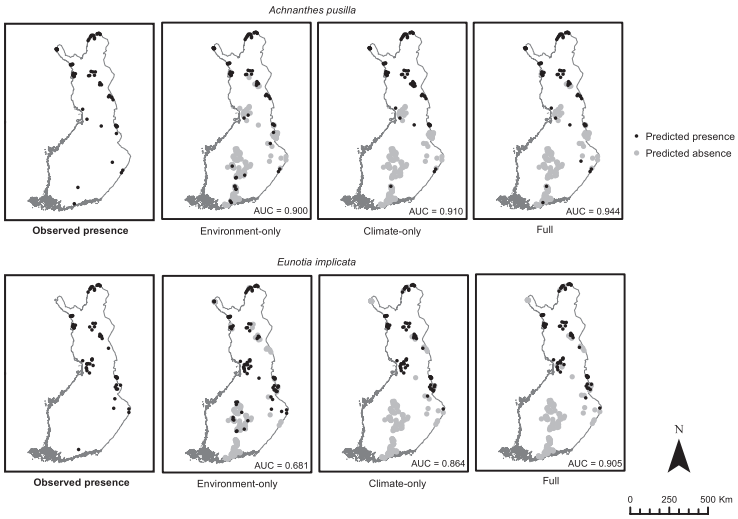


Figure 4 Observed and predicted distributions of two diatom species: *Achnanthes pusilla* and *Eunotia implicata*. The predicted distributions are modelled using the boosted regression tree method according to three predictor sets: environment only, climate only and the full set of predictors. The predictive performances of the models are evaluated using the area under the curve of a receiver operating characteristic plot (AUC).

DISCUSSION

We have shown here that climate is an important driver for diatom distributions. The climatic variables seem to have greater effect on stream diatom distributions than local environmental variables alone because climate-only models had a higher predictive performance than environment-only models. The importance of climate was further highlighted in variation partitioning and by the fact that climatic variables, namely GDD and PRECS, had the highest relative importance for diatom species distributions. The results of the species distribution models presented here suggest that the observed distribution patterns of stream diatoms can be explained reasonably well by climatic and local environmental predictors. Based on both AUC and TSS values the majority of the models had at least intermediate predictive performance and the model predictability was typical for freshwater organisms (Soininen & Luoto, 2014).

Our results support the idea that diatoms do respond strongly to large-scale climatic variables (Weckström *et al.*, 1997; Leira & Sabater, 2005), and therefore, are not controlled solely by local factors (Vyverman *et al.*, 2007; Verleyen *et al.*, 2009). However, we highlight that the predictive performance was highest in the full models and the joint effects in variation partitioning were relatively high, which indicates the importance of both local (e.g. conductivity) and large-scale climatic factors (temperature and precipitation) in species distributions, and reveals that the drivers of microbial distributions operate at multiple spatial scales.

The present results suggest that filtering of microorganisms to a local site takes place via both climate and local environmental factors in a nested fashion and microorganisms seem to have similar distinct regional pools of species as found for larger organisms (Martiny *et al.*, 2006; Vyverman *et al.*, 2007; Lindström & Langenheder, 2012). The novelty of our approach boils down to the fact that we were able to quantify the relative roles of climate versus local environmental variables for the distribution of individual taxa of microorganisms unlike most previous studies that have used community models (Potapova & Charles, 2002; Soininen *et al.*, 2004). Indeed, our method revealed broad among-species variation in predictability and variable importance even within a single taxonomic group, diatoms, suggesting that diatom species have unique responses towards local environment and climate. For example it has already been documented using weighted averaging that niche sizes of diatom species may vary greatly even within a single genus (Weckström *et al.*, 1997). However, we found no significant variation in predictability and variable importance among families. These results suggest that variability in predictability and variable importance exists at finer taxonomic resolution, whereas at coarser taxonomic resolution such variability is subtle. This may also indicate that occurrence data are somewhat noisy, thus increasing model uncertainty, which is reduced due to averaging on family level.

We emphasize the relatively modest spatial scale of our study and suggest that the climatic effect might have been even stronger if larger spatial scales, such as continents, had been

considered (Martiny *et al.*, 2006). This is because the relative importance of local environment and climatic factors in species occurrence may vary with study scale. It had been suggested previously that climatic effects may override any effects of local environmental factors at continental scales, while at regional scales (100–3000 km) microbial communities are influenced by both large-scale and local environmental factors (Martiny *et al.*, 2006; Astorga *et al.*, 2012). However, the present analyses suggest that climate also has a strong influence on diatom species distributions at smaller, regional scales. The higher predictive performance of climatic models also indicates that long-term climatic data, easily drawn from extensive databases, may be temporally more robust than water chemistry data based on snapshot field measurements. Nevertheless, although the variables used in present study did not show strong intercorrelations (all $r_s < |0.70|$) (Appendix S2), the possibility that climatic variables also reflect the influence of some latent local environmental variable (e.g. geology, catchment productivity) not included in the models cannot completely be ruled out with our study design. However, we suggest that climatic variables nonetheless provide excellent summary variables for modelling microbial distributions, although their effects are partly manifested via local variables.

Our study further reveals novel information about the importance of temperature (GDD) and precipitation for diatom distributions. GDD is a measure of heat accumulation and it influences primary production because temperature has a direct effect on metabolic processes (Atkinson, 1994; Brown *et al.*, 2004). Some of the diatom species are shown to have relatively narrow temperature ranges and a certain optimum water temperature (Weckström *et al.*, 1997). Our findings suggest, however, that air temperature sum during the growing season is a key predictor of diatom distribution. We emphasize that the effect of GDD on stream diatoms could also be indirect, as increasing GDD enhances the overall production rate in the catchment, including terrestrial productivity. In fact, recent studies have found that higher terrestrial catchment productivity was strongly associated with higher planktonic richness in lakes, possibly due to the elevated influx of carbon and inorganic nutrients to the surface water (Soininen & Luoto, 2012). Thus, we cannot tease apart the possible direct influence of GDD on diatom distributions from the indirect responses of diatoms to GDD via whole catchment productivity. A strong climatic effect was also highlighted by the importance of precipitation for diatom distributions. Summer precipitation can have an impact on riverine species via disturbance frequency (e.g. changes in current velocity, flooding, drought) (reviewed by Death, 2010). Current velocity is documented to have an impact on benthic diatom community composition even at small scales (Passy, 2001). Furthermore, precipitation causes runoff from the catchment areas, which serves as a source of nutrients and other substances in streams (Mallin *et al.*, 1993). It thus seems that energy (GDD) is a universal predictor of species distributions while precipitation may reflect both hydrology and chemical conditions via fluxes of elements to aquatic systems. These shared effects of climatic and local environmental factors highlight the complexity of environmental conditions faced by aquatic organisms, and therefore it is difficult to separate the pure effect of any single variable in observational studies. However, including both climatic and local environmental factors in SDM facilitates the understanding of the joint effects of climate and other environmental factors on species in future conditions.

In conclusion, our results suggest that large-scale climatic factors are important drivers of diatom distributions and they may be stronger than the local variables alone at relatively small, regional scales as well. Nevertheless, species distributions may be best predicted with both climatic and local environmental variables, highlighting their joint effects. These findings add to previous studies showing that microorganisms exhibit strong biogeographical patterns (Martiny *et al.*, 2006; Astorga *et al.*, 2012; Nemergut *et al.*, 2013). However, the predictability of the distributions of diatoms seems to vary greatly among species, suggesting that species have individualistic responses to environmental conditions and climate. Species-specific distribution modelling therefore emerges as an important new tool in explaining and predicting the responses of microorganisms to large-scale environmental gradients. We encourage researchers to further test the usefulness of SDMs for various microbial groups in different environments and using different spatial scales.

ACKNOWLEDGEMENTS

This project was funded by Maj and Tor Nessling foundation. We thank two anonymous referees for the insightful comments that greatly improved the manuscript.

REFERENCES

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Astorga, A., Oksanen, J., Luoto, M., Soininen, J., Virtanen, R. & Muotka, T. (2012) Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules? *Global Ecology and Biogeography*, **21**, 365–375.

Atkinson, D. (1994) Effects of temperature on the size of aquatic ectotherms: exceptions to the general rule. *Journal of Thermal Biology*, **20**, 61–74.

Battarbee, R.W. (2000) Palaeolimnological approaches to climate change, with special regard to the biological record. *Quaternary Science Reviews*, **19**, 107–124.

Bennett, J.R., Cumming, B.F., Ginn, B.K. & Smol, J.P. (2010) Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, **19**, 724–732.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M. & West, G.B. (2004) Toward a metabolic theory of ecology. *Ecology*, **85**, 1771–1789.

Crawley, M.J. (2007) Generalized linear models. *The R Book*, pp. 511–526. John Wiley and Sons, Chichester, UK.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. & Hess, K.T. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Davis, M.B. & Shaw, R.G. (2001) Range shifts and adaptive responses to Quaternary climate change. *Science*, **292**, 673–679.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243–251.

Death, R.G. (2010) Disturbance and riverine benthic communities: what has it contributed to general ecological theory? *River Research and Applications*, **26**, 15–25.

Elith, J., Graham, C., Anderson, R. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Eloranta, P. (1995) Type and quality of river waters in central Finland described using diatom indices. *Proceedings of the 13th International Diatom Symposium* (ed. by D. Marino and M. Montresor), pp. 271–280. Biopress, Bristol, UK.

Fielding, A.H. & Bell, J.F. (1997) A review methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Finlay, B. (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hastie, T. & Tibshirani, R. (1990) Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, **46**, 1005–1016.

Hijmans, R.J. (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, **93**, 679–688.

Holdridge, L.R. (1967) *Life zone ecology*. Tropical Science Center, Santa Jose, Costa Rica.

Kelly, M., Cazaubon, A., Coring, E. *et al.* (1998) Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology*, **10**, 215–224.

Korhonen, J.J., Köngäs, P. & Soininen, J. (2013) Temporal variation of diatom assemblages in oligotrophic and eutrophic streams. *European Journal of Phycology*, **48**, 141–151.

Krammer, K. & Lange-Bertalot, H. (1986–1991) *Bacillariophyceae. Süβwasserflora von Mitteleuropa 2 (1–4)*. Gustav Fischer Verlag, Stuttgart.

Ladau, J., Sharpton, T.J., Finucane, M.M., Jospin, G., Kembel, S.W., O'Dwyer, J., Koeppel, A.F., Green, J.L. & Pollard, K.S. (2013) Global marine bacterial diversity peaks at high latitudes in winter. *The ISME Journal*, **7**, 1669–1677.

Landis, J.R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lange-Bertalot, H. & Metzeltin, D. (1996) Indicators of oligotrophy – 800 taxa representative of three ecologically distinct lake types: carbonate buffered, oligodystrophic, weakly buffered soft water. *Iconographica diatomologica*, **2**, 1–390.

Leira, M. & Sabater, S. (2005) Diatom assemblages distribution in Catalan rivers, NE Spain, in relation to chemical and physiographical factors. *Water Research*, **39**, 73–82.

Levinsky, I., Araújo, M.B., Nogués-Bravo, D., Haywood, A.M., Valdes, P.J. & Rahbek, C. (2013) Climate envelope models suggest spatio-temporal co-occurrence of refugia of African birds and mammals. *Global Ecology and Biogeography*, **22**, 352–363.

Lindström, E.S. & Langenheder, S. (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**, 1–9.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–392.

Lugo, A., Brown, S., Dodson, R., Smith, T. & Shugart, H. (1999) The Holdridge life zones of the conterminous United States in relation to ecosystem mapping. *Journal of Biogeography*, **26**, 1025–1038.

Luoto, M., Virkkala, R. & Heikkinen, R.K. (2007) The role of land cover in bioclimatic models depends on spatial resolution. *Global Ecology and Biogeography*, **16**, 34–42.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*, 2nd edn. Chapman and Hall, New York.

Mallin, M.A., Pearl, H.W., Rudek, J. & Bates, P.W. (1993) Regulation of estuarine primary production by watershed rainfall and river flow. *Marine Ecology Progress Series*, **93**, 199–203.

Martiny, J.B.H., Bohannan, J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V.H. & Staley, J.T. (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews. Microbiology*, **4**, 102–112.

Nemergut, D.R., Schmidt, S.K., Fukami, T., O'Neill, S.P., Billinski, T.M., Stanish, L.F., Knelman, J.E., Darcy, J.L., Lynch, R.C., Wickey, P. & Ferrenberg, S. (2013) Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews*, **77**, 342–356.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H. & Wagner, H. (2015) *Vegan: Community Ecology Package*. Available at: http://cran.r-project.org/web/packages/vegan/index.html (accessed June 2015).

Passy, S.I. (2001) Spatial paradigms of lotic diatom distribution: a landscape ecology perspective. *Journal of Phycology*, **37**, 370–378.

Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

Potapova, M.G. & Charles, D.F. (2002) Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. *Journal of Biogeography*, **29**, 167–187.

R Development Core Team (2014) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Skov, F. & Svenning, J. (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography*, **27**, 366–380.

Soininen, J. & Luoto, M. (2012) Is catchment productivity a useful predictor of taxa richness in lake plankton communities? *Ecological Applications*, **22**, 624–633.

Soininen, J. & Luoto, M. (2014) Predictability in species distributions: a global analysis across organisms and ecosystems. *Global Ecology and Biogeography*, **23**, 1264–1274.

Soininen, J., Paavola, R. & Muotka, T. (2004) Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. *Ecography*, **27**, 330–342.

Soininen, J., Korhonen, J.J. & Luoto, M. (2013) Stochastic species distributions are driven by organism size. *Ecology*, **94**, 660–670.

Steinberg, C. (2003) *Ecology of humic substances in freshwaters: determinants from geochemistry to ecological niches*. Springer Science and Business Media, Berlin.

Swets, K. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Telford, R.J., Vandvik, V. & Birks, H.J.B. (2006) Dispersal limitations matter for microbial morphospecies. *Science*, **312**, 1015.

Thuiller, W., Albert, C., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, A., Hickler, T., Midgley, G.F., Paterson, J., Schurr, F.M., Sykes, M.T. & Zimmermann, N.E. (2008) Predicting global change impacts on plant species' distributions: future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, **9**, 137–152.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Vanormelingen, P., Verleyen, E. & Vyverman, W. (2008) The diversity and distribution of diatoms: from cosmopolitanism to narrow endemism. *Biodiversity and Conservation*, **17**, 393–405.

Venäläinen, A. & Heikinheimo, M. (2002) Meteorological data for agricultural applications. *Physics and Chemistry of the Earth*, **27**, 1045–1050.

Venier, L.A., Pearce, J., McKee, J.E., McKenney, D.W. & Niemi, G.J. (2004) Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Basin. *Journal of Biogeography*, **31**, 315–331.

Verleyen, E., Vyverman, W., Sterken, M., Hodgson, D.A., De Wever, A., Juggins, S., Van de Vijver, B., Jones, V.J., Vanormelingen, P., Roberts, D., Flower, R., Kilroy, C., Souffreau, C. & Sabbe, K. (2009) The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos*, **118**, 1239–1249.

Vyverman, W., Verleyen, E., Sabbe, K., Vanhoutte, K., Sterken, M., Hodgson, D.A., Mann, D.G., Juggins, S., Van de Vijver, B., Jones, V., Flower, R., Roberts, D., Chepurnov, V.A., Kilroy, C., Vanormelingen, P. & De Wever, A. (2007) Historical processes constrain patterns in global diatom diversity. *Ecology*, **88**, 1924–1931.

Walther, G.-R., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T.J.C., Fromentin, J.-M., Hoegh-Guldberg, O. & Bairlein, F. (2002) Ecological responses to recent climate change. *Nature*, **416**, 389–395.

Weckström, J., Korhola, A. & Blom, T. (1997) The relationship between diatoms and water temperature in thirty subarctic Fennoscandian lakes. *Arctic and Alpine Research*, **29**, 75–92.

Yee, T.W. & Mitchell, N.D. (1991) Generalized additive-models in plant ecology. *Journal of Vegetation Science*, **2**, 587–602.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Summary of environmental variables.
Appendix S2 The bivariate matrix for modelled variables.
Appendix S3 Wilcoxon signed rank tests for models.
Appendix S4 The predictive performance values for species distribution models.
Appendix S5 Variation partitioning for the species occurrence data among variable groups.

BIOSKETCHES

Virpi Pajunen is a PhD student in physical geography at the University of Helsinki. The focus of her thesis is in species distributions of benthic diatoms in streams.

Janne Soininen is an associate professor in spatial environmental research at the University of Helsinki. He is interested in large-scale community ecology, and especially in the distribution of small aquatic organisms.

Miska Luoto is a professor at the Department of Geosciences and Geography, University of Helsinki. His study interests are related to the integration of remote sensing and geographical information data in global change modelling.

Editor: Peter van Bodegom

Paper II

Pajunen, V., Jyrkänkallio-Mikkola, J., Luoto, M., Soininen, J. 2018. Are drivers of microbial bioindicators context dependent in human impacted and pristine environments? Submitted manuscript.

Are drivers of microbial bioindicators context dependent in human impacted and pristine environments?

Virpi Pajunen[1], Jenny Jyrkänkallio-Mikkola[1], Miska Luoto[1] and Janne Soininen[1]

[1]Department of Geosciences and Geography, P.O. Box 64, FI-00014 University of Helsinki, virpi.pajunen@helsinki.fi, jenny.jyrkankallio-mikkola@helsinki.fi, miska.luoto@helsinki.fi, janne.soininen@helsinki.fi

Abstract

Species occurrences are influenced by numerous factors of which effects may be context dependent. Thus, the magnitude of such effects and their relative importance on species distributions may vary among ecosystems due to anthropogenic stressors, for example. To investigate context dependency in factors governing microbial bioindicators, we developed species distribution models (SDMs) for stream diatom species separately in human impacted and pristine sites. We performed SDMs using boosted regression trees for 110 stream diatom species, which were common to both data sets, separately in 164 human impacted and 164 pristine sites in Finland (c. 1000 km, 60˚ – 68˚ N). For each species and site group, two sets of models were conducted: climate model, comprising three climatic variables, and full model, comprising the climatic and six local environmental variables. No significant difference in model performance was found between the site groups. However, climatic variables had a greater importance compared with local environmental variables in pristine sites, whereas local environmental variables had a greater importance in human impacted sites as hypothesized. Water balance and conductivity were the key variables in human impacted sites. The relative importance of climatic and local environmental variables varied among individual species, but also between the site groups. We found a clear context dependency among the variables influencing stream diatom distributions as the most important factors varied both among species and between the site groups. In human impacted streams, species distributions were mainly governed by water chemistry, whereas in pristine streams by climate. We suggest that climatic models may be suitable in pristine ecosystems, whereas the full models comprising both climatic and local environmental variables should be used in human impacted ecosystems.

Key words: climate; land use gradient; local environment; species distribution modelling; stream diatoms

1 Introduction

Ecosystems are molded by a myriad of factors operating at multiple spatial scales (Cox *et al.* 2016). This is especially true in open systems, such as rivers and streams, characterized by a unidirectional flow and a supply of substances from terrestrial areas (Allan and Castillo 2007). Large scale factors, such as climate and catchment land use, strongly influence the local stream habitat and thus the species diversity therein (Allan 2004, Pajunen *et al*. 2017). Due to the ongoing anthropogenic environmental change, streams are subjected to multiple stressors including changes in land use and associated habitat degradation and changing climatic conditions. This increasing stress is resulting in biodiversity loss and homogenization of communities (Rahel 2002, Olden *et al.* 2004, Filipe *et al*. 2013, Dar and Reshi 2014). As a consequence of global warming, the stream water temperatures are predicted to rise correspondingly (Webb 1996, Morrill *et al*. 2005) and changes in precipitation will alter hydrological conditions. The complex interactions between climate, land use and water physicochemistry challenge the future predictions of stream conditions.

The catchment land use affects stream physicochemistry (Foley *et al*. 2005) and even the past land use can have a long lasting imprint on stream conditions (Maloney *et al*. 2008, Walter and Merritts 2008, Maloney and Weller 2011).

For example, land cover previously dominated by agriculture may sustain high nutrient loads from sediments to streams for a long time period after a change in land use (Maloney and Weller 2011). Anthropogenic land use, comprising agriculture, development and urbanization, contributes to increased nutrient and ion concentrations (Taka *et al.* 2017), pollutants and turbidity (due to sediment load) in streams (Foley *et al.* 2005), and these effects cascade downstream (Levesque *et al.* 2017). Wang *et al.* (2008) found that nutrient loading and percentage of urban land use were the most important drivers of deteriorating stream conditions. Climate also affects nutrient levels in streams as nutrient fluxes in a stream network are strongly driven by hydrology (Arvola *et al.* 2015). Furthermore, riparian vegetation regulates stream temperature and light conditions by shading, acts as organic matter input and as an agent in sediment retention. Its removal leads to increased water temperature and sediment load (Studinski *et al*. 2012), but also to increased periphyton biomass due to greater light intensities (Von Schiller *et al.* 2007). Such effects of land use are likely to increase with projected higher air temperatures and precipitation in the future (e.g., Holmberg *et al*. 2006, Piggott *et al*. 2015). The relationship between land use and microbial stream communities strengthens towards downstream because of the continuous accumulation of substances in a river continuum (Tudesque *et al.* 2014). Moreover, stream microbes may show in some circumstances stronger relationship with the changes in land use than with physicochemical gradients, e.g., pH, substrates and nutrients (Liu *et al.* 2016, Jyrkänkallio-Mikkola *et al.* 2017). This indicates that land use could provide a more robust measure of water chemistry variables – thus, reflecting stream chemistry at longer time scales than snapshot water samples. This implies that especially in the presence of human activities, microbial communities are strongly influenced by the long gradients of local environmental factors (for example water chemistry) brought about by anthropogenic land use.

Biological indicators, such as benthic diatoms, are widely used to assess the ecological status of freshwater ecosystems as they reflect water quality over a period of time (Sandin and Verdonschot, 2006). Many aquatic microbes have species-specific responses towards water chemistry (Van Dam *et al.* 1994, Olapade and Leff 2005), but whether these responses stem from niche conservation or local adaption, is currently under debate (Finlay 2002, Wiens and Graham 2005). The relative importance of environmental variables (such as water chemistry and land use) affecting aquatic microbes can vary between study regions and in different climatic zones (Charles *et al.* 2006, Jüttner *et al.* 2010), and are also influenced by the study scale (Verleyen *et al.* 2009, Heino *et al.* 2014). This suggests a certain context dependency among the most influential factors driving microbial distributions. Furthermore, previous studies have shown a strong influence of climatic factors on the distributions of stream micro-organisms, which can even exceed the effect of local environmental variables (Pajunen *et al.* 2016). Climate can be seen as a crucial factor that has a strong impact on water temperatures and terrestrial vegetation patterns (Cox *et al*. 2016), and thus also on variation in in-stream variables (Frissell *et al*. 1986, Stevenson 1997). The effect of climate is likely to be more apparent in pristine environments where, in the absence of human impact, natural processes are able to dictate the supply of substances and the disturbance regime in streams. The gradients in water chemistry are expected to be shorter than in human impacted environments, thus the relative role of climatic factors affecting stream communities may be stronger. In contrast, in human impacted environments, local environment sets a strong filter for species due to the wide gradi-

ents in environmental factors.

To investigate whether the distributions of commonly used microbial bioindicators are context dependent, i.e. species' responses vary between species and among sites with different magnitude of human impact, we developed species distribution models (SDMs) for stream diatoms separately in human impacted and pristine streams. Earlier studies have mostly concentrated on changes in community composition between different gradients of human impact (e.g., Pan *et al.* 2004, Soininen *et al.* 2004, Hering *et al.* 2006), yet the knowledge about the responses of individual species to environmental and climatic factors in different environments are still scarce. We hypothesized that climatic variables affect the distribution of diatom species more in pristine sites than in human impacted sites. As a corollary, the effect of local environmental variables on species distributions is stronger in human impacted sites where the gradients of water chemistry are longer than in pristine sites. In human impacted sites, the addition of local variables to climate models would thus greatly enhance the model performance.

2 Methods

2.1 Data sampling and analysis

The data set comprised diatom (presence/absence), water chemistry and physical variable data collected from Finnish stream sites between 1986 and 2016 (328 sites in total) (Fig. 1). The samples were considered comparable as the sampling methods were identical and all sampling was performed during the base flow conditions in July to September. The sites were distributed relatively evenly across Finland and the measured environmental and climatic variables covered a wide gradient (Appendix S1: Table S1). More detailed information about the data set can be found in Eloranta (1995), Soininen *et al.* (2004) and Jyrkänkallio-Mikkola *et al.* (2017).

Each stream site was sampled for diatoms by collecting five to ten replicate cobble sized stones. Biofilm was removed from the stones by brushing them with a toothbrush. Water samples were taken simultaneously with diatom samples and were subsequently analyzed for total phosphorus (TP), pH, conductivity and water color using national standards. For the minority of the sites (< 10 %), water chemistry data were taken from the national water quality database, using results from the nearest sampling occasion. Current velocity, canopy shading and stream width were measured at each site along the site perpendicular to the flow and covering the whole riffle. Samples were cleaned from organic material in the laboratory using wet combustion with acid (HNO3:H2SO4; 2:1 or hydrogen peroxide [30%, H2O2]) and mounted in Naphrax or Dirax. A total of 250–500 diatom frustules per sample were identified to the lowest possible taxonomic level according to Krammer and Lange-Bertalot (1986–1991) and Lange-Bertalot and Metzeltin (1996), and counted using phase contrast light microscopy (magnification 1000×). A species was considered present at a site when at least one valve was observed.

2.2 Climatic and land use variables

We compiled a set of environmental variables presumed to affect diatom distributions. Climatic variables were chosen based on their use in previous SDMs conducted for Finnish diatom data (Pajunen *et al.* 2016). The variables were growing degree days adjusted to 5 ˚C (GDD), season precipitation sum from May to September (PRECS) and water balance (WAB; calculated according to Skov and Svenning 2004). GDD represents the aerial temperature and the energy

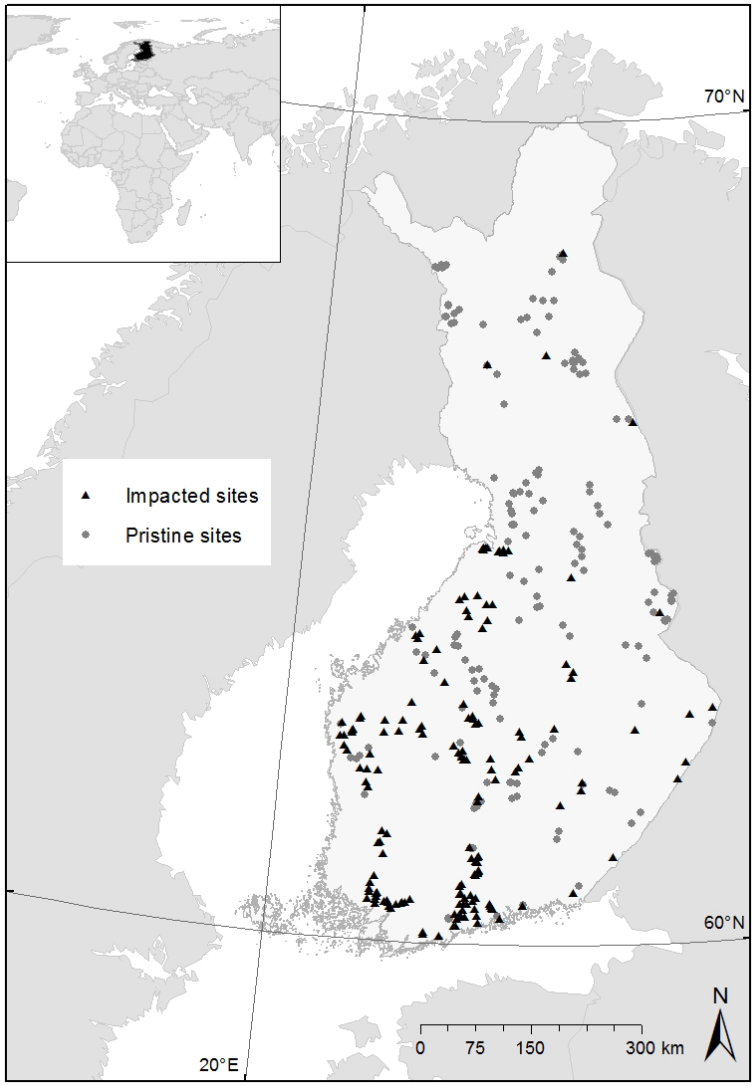


Figure 1 Location of the sampling sites (n = 328) in Finland, northern Europe, divided into two subgroups: human impacted sites (n = 164, > 5% anthropogenic land use) and pristine sites (n = 164, < 5% anthropogenic land use). The index map represents the location of Finland in the Northern Hemisphere.

requirements of the species, while PRECS and WAB represent the moisture availability in the environment, connected to the extent of recharge and run-off. The climatic data set covered the years 1981–2010 and was obtained as a 10×10 km resolution grid from the Finnish Meteorological Institute (Venäläinen and Heikinheimo 2002). Using ArcGIS 10.3.1 software, site-specific catchment areas were created by calculating the patterns of flow direction and accumulation to each sampling point from digital elevation model (DEM; grid resolution 10×10 m, National Land Survey of Finland 2013). Classifications of land use were obtained from CORINE Land Cover data (20×20 m, Finnish Environment Institute 2013). Artificial and agricultural land use were merged to represent anthropogenic land use. The local and climatic variables were tested for co-

variance with nonparametric Spearman's rank correlation coefficient. All predictor variables had low collinearity ($r_s \leq |0.50|$, Appendix S1: Figs. S1–2).

2.3 Species distribution models

The data set (328 sites and in total 494 diatom species) was divided into two equal-sized groups: human impacted sites (n = 164, > 5% anthropogenic land use) and pristine sites (n = 164, < 5% anthropogenic land use). The 5% threshold for anthropogenic land use fits the data set as Finland consists mainly of forested areas and scattered settlement. Therefore, even a small increase in human impact can have a notable effect on stream conditions. Diatom species that occurred in both groups and at least at 5% and maximum at 95% of the sites were included in the statistical analyses. Two sets of diatom species distribution models were conducted for each of the 110 species separately for human impacted and pristine sites: climate models and full models. In the climate models, species distributions were modelled only by the three climatic variables: GDD, PRECS and WAB. In addition to the climatic variables, the full models had six environmental predictors: TP, conductivity, pH, water color (mainly reflecting the humic content of the water), canopy shading and current velocity.

The SDMs were applied via the BIOMOD2 framework (Thuiller *et al.* 2016) fitted in R (version 3.3.3; R Development Core Team 2017) using boosted regression trees (BRT) as the modelling algorithm. BRT is a machine learning technique, which has previously proven to be a robust method for creating SDMs for micro-organisms (Pajunen *et al.* 2016), as it is highly efficient at fitting nonparametric data, and can manage various types of predictor variables. It does not require prior data transformation and takes automatically into account the interaction effects between predictors (the principles of BRT in more detail: see Friedman 2001, De'ath 2007, Elith *et al.* 2008). BRTs were performed with a maximum number of 3000 trees, the interaction depth of 6 and the learning rate of 0.001.

The performance of each model was assessed with a cross validation (CV) approach, where the models were fitted four times by using a random sample of 70% of the data and subsequently evaluated against the remaining 30%. The predicted and observed occurrences of species were compared at each CV run by calculating the area under the curve of a receiver operating characteristic plot (AUC) (Fielding and Bell 1997) and true skill statistics (TSS) (Allouche *et al.* 2006). The models have at least intermediate predictive performance if AUC values are > 0.7 (following Swets 1998) and TSS values are > 0.4 (following Landis and Koch 1977).

The importance of each predictor for a species in the models was assessed in BIOMOD2 by randomizing each variable individually and then projecting the model with the randomized variable while keeping the other variables unchanged. The model predictions containing the randomized variable were further correlated with those of the original models. Finally, the importance of the variable was calculated as one minus the correlation; higher values indicate predictors that are more important for the model (Thuiller *et al.* 2009). This analysis was repeated ten times. The differences in model performances and predictor relative importances between human impacted and pristine data sets were tested using a paired t-test.

3 Results

Both climatic and full models performed satisfactorily in human impacted and pristine sites (SDM averages AUC > 0.70 and TSS > 0.40) and

all sets of SDMs (i.e. climate and full models in both site groups) had similar patterns in predictive performances (Appendix S2: Fig. S1). The inclusion of local environmental variables into the models did not improve the model performance compared to the climate models in pristine sites (climate model AUC 0.710 and TSS 0.447; full model AUC 0.707 and TSS 0.435), whereas in human impacted sites it slightly did (climate model AUC 0.708 and TSS 0.442; full model AUC 0.725 and TSS 0.464). However, no significant difference in model performance was found between human impacted and pristine sites (paired t-test, all $P > 0.05$).

Agreeing with our hypothesis, in the full models, climatic variables had on average greater variable importance (the sum of median (md) importance: climatic 55% and local environmen-

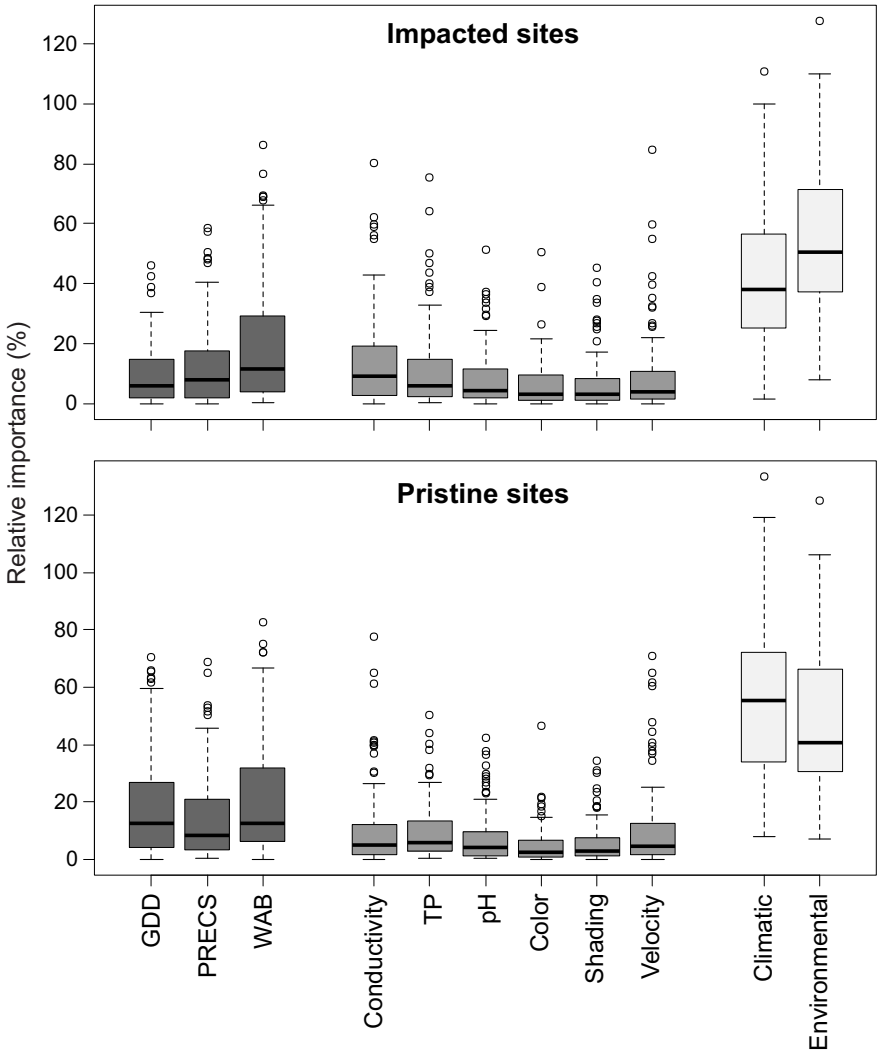


Figure 2 Relative importance (%) of climatic and local environmental variables and the sums of both variable groups for diatom species (n = 110) distributions separately in human impacted sites (n = 164, > 5% anthropogenic land use) and pristine sites (n = 164, < 5% anthropogenic land use). Models were conducted using boosted regression trees and the full set of predictors. The abbreviations stand for growing degree days (GDD), precipitation (PRECS), water balance (WAB) and total phosphorus (TP). Error bars represent standard errors.

tal variables 41%) compared with local environmental variables in pristine sites. In human impacted sites local environmental variables were more important (climatic 38% and local environmental variables 50%, respectively) (Fig. 2). WAB was the most important variable in both site groups (md importance: human impacted 11% and pristine sites 12%), whereas GDD was as important in pristine sites (md = 12%). In human impacted sites, conductivity had the second greatest relative importance (md = 9%) on species distributions. PRECS had the second greatest importance (md = 8%) in pristine sites and the third greatest (md = 8%) in human impacted sites. TP had the third greatest importance (md = 6%) on species distributions in pristine sites.

The relative importance of climatic and local environmental variables on individual species varied among species, but also between human impacted and pristine sites (Fig. 3, Appendix S2: Fig. S2). Overall variation between the two site groups was significant only for the relative importance of GDD (paired t-test; $P < 0.001$) being higher in pristine sites, whereas for other variables it was nonsignificant (paired t-test; $P > 0.05$). The between-site group variation among the relative importance of variables on individual species was species-specific: some species responded similarly to climatic and/or local environmental variables in both site groups, yet the responses of some other species varied greatly (e.g. *Cocconeis placentula*, *Navicula rhynchocephala*) (Appendix S2: Fig. S3).

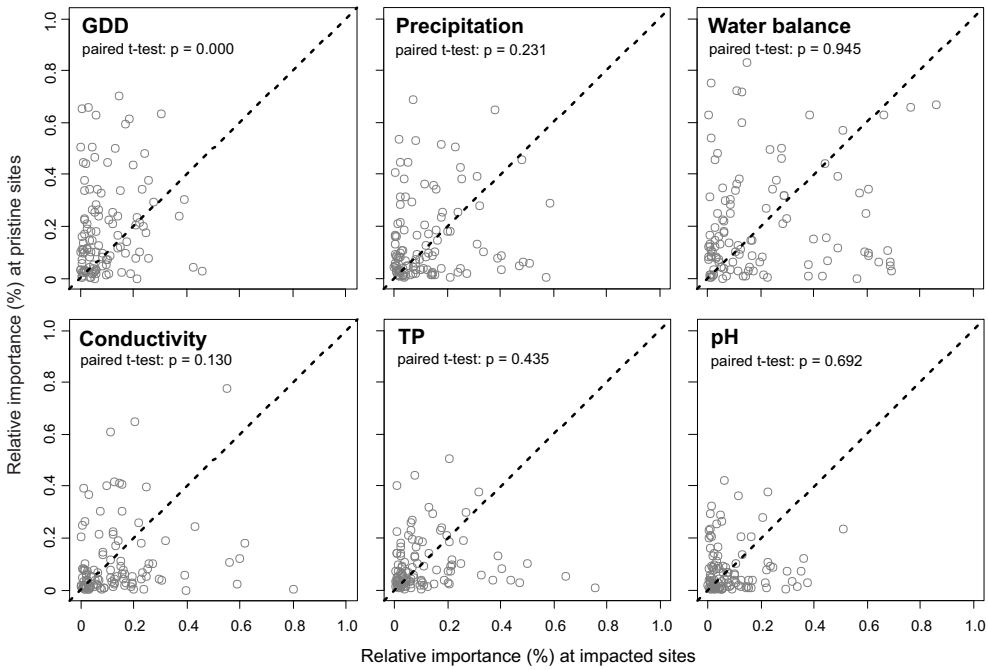


Figure 3 Relationships between the relative importance of six predictors for diatom species distribution in human impacted and pristine sites. The models were conducted using boosted regression trees as modelling method and the full set of predictors. The full models consist of three climatic predictors (growing degree days (GDD), precipitation (PRECS) and water balance (WAB)) and six local environmental predictors (conductivity, total phosphorus (TP), pH, water color, shading by the canopy and current velocity). The differences between the site groups are compared in each plot using paired t-test. Dashed lines demonstrate the diagonal line (0, 1).

4 Discussion

Our study reveals a notable context dependency among the factors influencing the distributions of diatom species. The most important factors affecting diatom species distribution vary not only among species, but are also dependent on the degree of anthropogenic influence. Consistent with our hypothesis, the overall importance of climatic variables on species occurrence was greater in pristine streams than was the importance of local environmental variables. In contrast, the importance of local environmental variables was the greatest in human impacted sites. However, we emphasize that water balance has a significant impact on stream diatom distributions in all stream environments, and its effect was stronger than the influence of any single local environmental variable not only in pristine locations, but also in human impacted sites. This corresponds to previous studies indicating that diatom species can be shaped by large-scale climatic and historical factors (Weckström *et al.* 1997a, Leira and Sabater 2005, Vyverman *et al.* 2007, Pajunen *et al.* 2016).

Although climate is the ultimate factor influencing stream diatom distributions both directly (temperature) and indirectly (productivity and hydrology) (Pajunen *et al.* 2016, 2017), the importance of local physicochemical factors seems to be highlighted in the streams influenced by anthropogenic activities. This can be partly explained by the wide gradients of water chemistry variables (here conductivity and TP; see Appendix S1: Table S1) related to anthropogenic land use and by the strong species responses towards these variables (i.e. species filtering along environmental gradients). The tolerances of individual diatom species towards local environmental factors have been widely studied and many species have restricted tolerances and preferences towards certain environmental variables (for example, nutrients [Winter and Duthie 2000], conductivity [Potapova and Charles 2003] and pH [Andrén and Jarlman 2008]). For example, diatom communities in streams under human impact, such as agriculture and point-source pollution, consist of species with a preference to high nutrient levels and tolerance towards pollutants (Lavoie *et al*. 2006, Moravcova *et al*. 2013).

Recently, growing evidence of context dependency in species responses toward environmental and spatial factors has been documented among stream organisms (for example, diatoms [Heino *et al*. 2012], bryophytes [Heino *et al.* 2012] and macroinvertebrates [Heino *et al.* 2012, Hawkins *et al.* 2015, Tonkin *et al*. 2016]). Stream diatoms are simultaneously affected by abiotic and biotic forcing whose relative importance differs among sites and regions (Clements *et al.* 2015). For instance, the occurrence of *Frustulia rhomboides*, a species often classified as acidophilous i.e. occurring at pH <7 (Van Dam *et al.* 1994, Weckström *et al*. 1997b), was affected mainly by pH in pristine sites (relative importance = 32%). However, in human impacted sites, it was mostly affected by conductivity (relative importance = 60%), while the relative importance of pH was negligible (1%, see Appendix S2: Fig. S3). Among-region variation in species-specific and community-level responses of diatoms to water chemistry has also been observed elsewhere (Charles *et al.* 2006, Jüttner *et al.* 2010, Chen *et al.* 2016). For example, Chen *et al.* (2016) found that diatom species indicating high nutrient conditions in U.S. streams occurred in low nutrient streams in China suggesting that diatom niches were not conserved. However, studies from different regions should be compared with caution as morphological taxonomic identification may contain locally adapted morphotypes of species (Rose and Cox 2014). Also, the spatial scale of the study may affect the rela-

tive importance of the most influential factors. The effect of climatic factors operating at large spatial scales may become more important when the spatial scale is large (Martiny *et al.* 2006).

Context dependency may also be a sign of genotypic plasticity, i.e. species can be adapted to local conditions through rapid genetic evolution, enabled by the fast life-cycle of microbes (Birch 1960). Or, it may reflect phenotypic plasticity, i.e. the tolerances towards environmental factors vary among different morphotypes of individual species, which results in variable responses (Rose and Cox 2014). However, there is also clear evidence for niche conservation in lacustrine diatoms (Bennett *et al.* 2010). Additionally, the species' responses along anthropogenic gradients may also vary due to biotic interactions (such as competition), which intensity may vary along the shifts in community structure (Tilman 1977, Stelzer and Lamberti 2001). The number of possible processes causing context dependency highlights the need to study this topic further in the near future.

The moisture related factors, i.e. WAB and precipitation, were important both in human impacted and pristine sites. Precipitation and run-off are essential factors influencing aquatic biota, including stream diatoms, via weathering, transport of substances and the flow regime (Stevenson *et al.* 1996, Leland and Porter 2000, Allan and Castillo 2007). In human impacted streams, these climatic variables can enhance the effect of anthropogenic land use through run-off, which can consist of high amounts of allochthonous nutrients, organic matter, solids and pollutants (Pan *et al.* 2004, Death *et al.* 2015, Ponsati *et al.* 2016). The importance of climate is further emphasized by the fact that the impact of land use on water chemistry and further on diatom communities may weaken during summer base-flow conditions compared to wetter seasons (Pan *et al.* 2004). The hierarchical structure of environmental factors (for instance, climate influencing land cover which affects water physicochemistry) (Frissell *et al.* 1986, Stevenson 1997) may become more evident in the absence of human impact. For example, the relative importance of GDD was significantly greater in pristine than in human impacted sites suggesting that both the direct (temperature) and indirect (for example catchment and in-stream productivity) effects of GDD are more essential drivers in more pristine systems (Fig. 3). However, the relative importance of climatic variables in the human impacted data set may be influenced by the fact that the anthropogenic land use is mostly situated in the southern and western regions of Finland where the growing season is the longest. Thus, the human impacted data set contains less sites with cold and dry climatic conditions, more typical in the northern regions, compared to the pristine data set.

In conclusion, we found that the main drivers of diatom species distributions and also the species-specific responses to these drivers differed among human impacted and pristine environments. The effect of climate was important both in pristine and in human impacted streams in spite of wide gradients in local environmental variables and anthropogenic land use in the latter. However, the climatic influence was the strongest in pristine streams, suggesting that climatic variables need to be considered in diatom models especially in regions where water chemistry gradients are only modest and stream physicochemistry is mainly dictated by natural landscape and the processes therein. The way that climatic and environmental change will alter stream conditions in the future may be context dependent and differ among environments. Thus, it will be challenging to predict the distribution of microorganisms under future climate scenarios.

Acknowledgements

This project was funded by Maj and Tor Nessling foundation, Nordenskiöld foundation and the Academy of Finland (grant 273560).

References

Allan, J. D. 2004. Landscapes and riverscapes: the influence of land use on stream ecosystems. Annual Review of Ecology and Systematics 35:257–284.

Allan, J. D., and M. M. Castillo. 2007. Stream Ecology: Structure and Function on Running Waters. 2nd edn. Springer, Dordrecht, the Netherlands.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43:1223–1232.

Andrén, C., and A. Jarlman. 2008. Benthic diatoms as indicators of acidity in streams. Fundamental and Applied Limnology 173:237–253.

Arvola, L., E. Einola, and M. Järvinen. 2015. Landscape properties and precipitation as determinants for high summer nitrogen load from boreal catchments. Landscape Ecology 30:429–442.

Bennett, J. R., B. F. Cumming, B. K. Ginn, and J. P. Smol. 2010. Broad-scale environmental response and niche conservatism in lacustrine diatom communities. Global Ecology and Biogeography 19:724–732.

Birch, L. C. 1960. The genetic factor in population ecology. The American Naturalist 94:5–24.

Charles, D. F., F. W. Acker, D. D. Hart, C. W. Reimer, and P. B. Cotter. 2006. Large-scale regional variation in diatom-water chemistry relationships: Rivers of the eastern United States. Hydrobiologia 561:27–57.

Chen, X., W. Zhou, S. T. A. Pickett, W. Li, L. Han, and Y. Ren. 2016. Diatoms are better indicators of urban stream conditions: A case study in Beijing, China. Ecological Indicators 60:265–274.

Clements, W. H., D. R. Kashian, P. M. Kiffney, and R. E. Zuellig. 2016. Perspectives on the context-dependency of stream community responses to contaminants. Freshwater Biology 61:2162–2170.

Cox, C. B., P. D. Moore, and R. J. Ladle. 2016. Biogeography: An Ecological and Evolutionary Approach. John Wiley and Sons Ltd, Chichester, UK.

Dar, P. A., and Z. A. Reshi. 2014. Components, processes and consequences of biotic homogenization: A review. Contemporary Problems of Ecology 7:123–136.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. Ecology 88:243–251.

Death, R. G., I. C. Fuller, and M. G. Macklin. 2015. Resetting the river template: the potential for climate-related extreme floods to transform river geomorphology and ecology. Freshwater Biology 60:2477–2496.

Elith J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

Eloranta, P. 1995. Type and quality of river waters in central Finland described using diatom indices. Pages 271–280 in D. Marino, and M. Montresor, editors. Proceedings of the 13th International Diatom Symposium. Biopress, Bristol, UK.

Fielding, A. H., and J. F. Bell. 1997. A review methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.

Filipe, A. F., J. E. Lawrence, and N. Bonada. 2013. Vulnerability of stream biota to climate change in mediterranean climate regions: a synthesis of ecological responses and conservation challenges. Hydrobiologia 719:331–351.

Finlay, B. 2002. Global dispersal of free-living microbial eukaryote species. Science 296:1061–1063.

Finnish Environment Institute. 2013. CORINE Land Cover 20 m. https://avaa.tdata.fi/web/paituli. Accessed September 14, 2017.

Foley, J. A., R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P. K. Snyder. 2005. Global consequences of land use. Science 309:570–574.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29:1189–1232.

Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. Environmental Management 10:199-214.

Hawkins, C. P., H. Mykrä, J. Oksanen, and J. J. Varder Laan. 2015. Environmental disturbance can increase beta diversity of stream macroinvertebrate assemblages. Global Ecology and Biogeography 24:483–494.

Heino, J., M. Grönroos, J. Soininen, R. Virtanen, and T. Muotka. 2012. Context dependency and metacommunity structuring in boreal headwater streams. Oikos 121:537–544.

Heino, J., M. Tolkkinen, A. M. Pirttilä, H. Aisala, and H. Mykrä. 2014. Microbial diversity and community-environment relationships in boreal streams. Journal of Biogeography 41:2234–2244.

Hering, D., R. K. Johnson, S. Kramm, S. Schmutz, K. Szoszkiewicz, and P. F. M. Verdonschot. 2006. Assessment of European streams with diatoms, macrophytes, macroinvertebrates and fish: a comparative metric-based analysis of organism response to stress. Freshwater Biology 51:1757–1785.

Holmberg, M., M. Forsius, M. Starr, and M. Huttunen. 2006. An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change. Ecological Modelling 195:51–60.

Jüttner, I., P. D. J. Chimonides, S. J. Ormerod, and E. J. Cox. 2010. Ecology and biogeography of Himalayan diatoms: distribution along gradients of altitude, stream habitat and water chemistry. Fundamental and Applied Limnology 177:293–311.

Jyrkänkallio-Mikkola, J., S. Meier, J. Heino, T. Laamanen, V. Pajunen, K. T. Tolonen, M. Tolkkinen, and J. Soininen. 2017. Disentangling multi-scale environmental effects on stream microbial communities. Journal of Biogeography 44:1512–1523.

Krammer, K., and H. Lange-Bertalot. 1986–1991. Bacillariophyceae. Süßwasserflora von Mitteleuropa 2 (1-4). Gustav Fischer Verlag, Stuttgart, Germany.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159–174.

Lange-Bertalot, H., and D. Metzeltin. 1996. Iconographica diatomologica, Volume 2. Indicators of oligotrophy. 800 taxa representative of three ecologically distinct lake types: carbonate buffered, oligodystrophic, weakly buffered soft water. Koeltz Scientific Books, Koenigstein, Germany.

Lavoie, I., S. Campeau, M. Grenier, and P. J. Dillon. 2006. A diatom-based index for the biological assessment of eastern Canadian rivers: an application of correspondence analysis (CA). Canadian Journal of Fisheries and Aquatic Sciences 63:1793–1811.

Leira M., and S. Sabater. 2005. Diatom assemblages distribution in catalan rivers, NE Spain, in relation to chemical and physiographical factors. Water research 39:73–82.

Levesque, D., C. Hudon, P. M. A. James, and P. Legendre. 2017. Environmental factors structuring benthic primary producers at different spatial scales in the St. Lawrence River (Canada). Aquatic Sciences 79:345–356.

Liu, S., G. Xie, L. Wang, K. Cottenie, D. Liu, and B. Wang. 2016. Different roles of environmental variables and spatial factors in structuring stream benthic diatom and macroinvertebrate in Yangtze River Delta, China. Ecological Indicators 61:602–611.

Maloney, K. O., J. W. Feminella, R. M. Mitchell, S. A. Miller, P. J. Mulholland, and J. N. Houser. 2008. Landuse legacies and small streams: identifying relationships between historical land use and contemporary stream conditions. Journal of the North American Benthological Society 27:280–294.

Maloney, K. O., and D. E. Weller. 2011. Anthropogenic disturbance and streams: land use and land-use change affect stream ecosystems via multiple pathways. Freshwater Biology 56:611–626.

Martiny, J. B. H., J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Øvreås, A.-L. Reysenbach, V. H. Smith, and J. T. Staley. 2006. Microbial biogeography: putting microorganisms on the map. Nature Reviews Microbiology 4:102–112.

Moravcova, A., O. Rauch, J. Lukavsky, and L. Nedbalova. 2013. The responses of epilithic diatom assemblages to sewage pollution in mountain streams of the Czech Republic. Plant Ecology and Evolution 146:153–166.

Morrill, J. C., R. C. Bales, and M. H. Conklin. 2005. Estimating stream temperature from air temperature: Implications for future water quality. Journal of Environmental Engineering-asce 131:139–146.

National Land Survey of Finland. 2013. Elevation model 10 m. https://avaa.tdata.fi/web/paituli. Accessed September 14, 2017.

Olapade, O. A., and L. G. Leff. 2005. Seasonal response of stream biofilm communities to dissolved organic matter and nutrient enrichment. Applied and Environmental Microbiology 71:2278–2287.

Olden, J. D., N. L. Poff, M. R. Douglas, M. E. Douglas, and K. D. Fausch. 2004. Ecological and evolutionary consequences of biotic homogenization. Trends in Ecology & Evolution 19:18–24.

Pan, Y., A. Herlihy, P. Kaufmann, J. Wigington, J. van Sickle, and T. Moser. 2004. Linkages among land-use, water quality, physical habitat conditions and lotic diatom assemblages: A multi-spatial scale assessment. Hydrobiologia 515:59–73.

Pajunen, V., M. Luoto, and J. Soininen. 2016. Climate is an important driver for stream diatom distributions. Global Ecology and Biogeography 25:198–206.

Pajunen, V., M. Luoto, and J. Soininen. 2017. Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities. Journal of Biogeography 44:2376–2385.

Piggott, J. J., R. K. Salis, G. Lear, C. R. Townsend, and C. D. Matthaei. 2015. Climate warming and agricultural stressors interact to determine stream periphyton community composition. Global Change Biology 21:206–222.

Ponsatí, L., N. Corcoll, M. Petrovic, Y. Picó, A. Ginebreda, E. Tornés, H. Guash, D. Barceló, and S. Sabater. 2016. Multiple-stressor effects on river biofilms under different hydrological conditions. Freshwater Biology 61:2102–2115.

Potapova, M., and D. F. Charles. 2003. Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. Freshwater Biology 48:1311–1328.

R Development Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org. Accessed September 14, 2017.

Rahel, F. J. 2002. Homogenization of freshwater faunas. Annual Review of Ecology and Systematics 33:291–315.

Rose D. T., and E. J. Cox. 2014. What constitutes Gomphonema parvulum? Long-term culture studies show that some varieties of G. parvulum belong with other Gomphonema species. Plant Ecology and Evolution 147:366–373.

Sandin, L., and P. F. M. Verdonschot. 2006. Stream and river typologies – major results and conclusions from the STAR project. Hydrobiologia 566:33–37.

Skov, F., and J. Svenning. 2004. Potential impact of climatic change on the distribution of forest herbs in Europe. Ecography 27:366–380.

Soininen, J., R. Paavola, and T. Muotka. 2004. Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. Ecography 27:330–342.

Stelzer, R. S., and G. A. Lamberti. 2001. Effects of N : P ratio and total nutrient concentration on stream periphyton community structure, biomass, and elemental composition. Limnology and Oceanography 46:356–367.

Stevenson, R. J., M. L. Bothwell, and R. L. Lowe. 1996. Algal Ecology: Freshwater Benthic Ecosystems. Elsevier, San Diego, USA.

Stevenson, R. J. 1997. Scale-dependent determinants and consequences of benthic algal heterogeneity. Journal of North American Benthological Society 16:248–262.

Studinski, J. M., K. J. Hartman, J. M. Niles, and P. Keyser. 2012. The effects of riparian forest disturbance on stream temperature, sedimentation and morphology. Hydrobiologia 686:107–117.

Swets, K. 1988. Measuring the accuracy of diagnostic systems. Science 240:1285–1293.

Taka, M., T. Kokkonen,, K. Kuoppamäki, T. Niemi, N. Sillanpää, M. Valtanen, L. Warsta, and H. Setälä. 2017. Spatio-temporal patterns of major ions in urban stormwater under cold climate. Hydrological Processes 31:1564–1577.

Thuiller, W., D. Georges, R. Engler, and F. Breiner. 2016. biomod2: Ensemble Platform for Species Distribution Modeling. mran.microsoft.com/package/biomod2/biomod2.pdf. Accessed September 14, 2017.

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32:369–373.

Tilman, D. 1977. Resource competition between plankton algae: an experimental and theoretical approach. Ecology 58:338–348.

Tonkin, J. D., J. Heino, A. Sundermann, P. Haase, and S. C. Jähnig. 2016. Context dependency in biodiversity patterns of central German stream metacommunities. Freshwater Biology 61:607–620.

Tudesque, L., C. Tisseuil, and S. Lek. 2014. Scale-dependent effects of land cover on water physicochemistry and diatom-based metrics in a major river system, the Adour-Garonne basin (South Westerns France). Science of the Total Environment 466:47–55.

Van Dam, H., A. Mertens, and J. Sinkeldam. 1994. A coded checklist and ecological indicators values of freshwater diatoms from the Netherlands. Netherlands Journal of Aquatic Ecology 28:117–133.

Venäläinen, A., and M. Heikinheimo. 2002. Meteorological data for agricultural applications. Physics and Chemistry of the Earth 27:1045–1050.

Verleyen, E., W. Vyverman, M. Sterken, D. A. Hodgson, A. De Wever, S. Juggins, B. Van de Vijver, V. J. Jones, P. Vanormelingen, D. Roberts, R. Flower, C. Kilroy, C. Souffreau, and K. Sabbe. 2009. The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. Oikos 118:1239–1249.

Von Schiller, D., E. Marti, J. L. Riera, and F. Sabater. 2007. Effects of nutrients and light on periphyton biomass and nitrogen uptake in Mediterranean streams with contrasting land uses. Freshwater Biology 52:891–906.

Vyverman, W., E. Verleyen, K. Sabbe, K. Vanhoutte, M. Sterken, D. A. Hodgson, D. G. Mann, S. Juggins, B. Van de Vijver, V. Jones, R. Flower, D. Roberts, V. A. Chepurnov, C. Kilroy, P. Vanormelingen, and A. De Wever. 2007. Historical processes constrain patterns in global diatom diversitymen. Ecology 88:1924–1931.

Walter, R. C., and D. J. Merritts. 2008. Natural streams and the legacy of water-powered mills. Science 319:299–304.

Wang, L., T. Brenden, P. Seelbach, A. Cooper, D. Allan, R. Clark Jr., and M. Wiley. 2008. Landscape based identification of human disturbance gradients and reference conditions for Michigan streams. Environmental Monitoring and Assessment 141:1–17.

Webb, B. W. 1996. Trends in stream and river temperature. Hydrological Processes 10:205–226.

Weckström, J., A. Korhola, and T. Blom. 1997a. The relationship between diatoms and water temperature in thirty subarctic Fennoscandian lakes. Arctic and Alpine Research 29:75–92.

Weckström, J., A. Korhola, and T. Blom. 1997b. Diatoms as quantitative indicators of pH and water temperature in subarctic Fennoscandian lakes. Hydrobiologia 347:171–184.

Wiens, J. J., and C. H. Graham. 2005. Niche conservatism: Integrating evolution, ecology, and conservation biology. Annual Review of Ecology, Evolution, and Systematics 36:519–539.

Winter, J., and H. Duthie. 2000. Epilithic diatoms as indicators of stream total N and total P concentration. Journal of the North American Benthological Society 19:32–49.

APPENDIX S1 Summary and correlation of variables.

Table S1 The summary (minimum, maximum, range, median and standard deviation (Sd)) of the measured variables from 164 impacted (> 5% anthropogenic land use) and 164 pristine (< 5% anthropogenic land use) stream sites in Finland.

		Impacted sites					**Pristine sites**				
Variable	Unit	Min	Max	Range	Median	Sd	Min	Max	Range	Median	Sd
Growing degree days		631.0	1465.9	834.8	1193.1	130.5	531.7	1450.6	919.0	953.8	220.5
Precipitation	mm	273.2	343.1	69.9	310.4	12.2	253.8	346.9	93.1	313.7	25.5
Water balance	mm	269.9	573.4	303.5	364.0	63.5	272.8	624.1	351.3	332.0	95.7
Total phosphorus	$\mu g\ L^{-1}$	2.0	356.9	354.9	48.5	49.4	0.1	182.5	182.4	18.3	24.9
Conductivity	$\mu S\ cm^{-1}$	12.9	619.0	606.1	91.6	88.9	9.4	161.1	151.7	30	24.3
pH		5.7	8.2	2.5	7.1	0.5	4.5	8.1	3.6	6.7	0.6
Water color	$mg\ Pt\ L^{-1}$	5.0	375.0	370.0	90.0	74.9	2.5	625	622.5	100	87.0
Shading	%	0.0	100.0	100.0	39.8	26.4	0	100	100	35.5	26.0
Current velocity	$m\ s^{-1}$	0.0	1.7	1.6	0.3	0.3	0.0	1.7	1.7	0.3	0.3
Anthropogenic land use	%	5.0	65.7	60.8	18.5	15.0	0.0	4.9	4.9	1.3	1.5

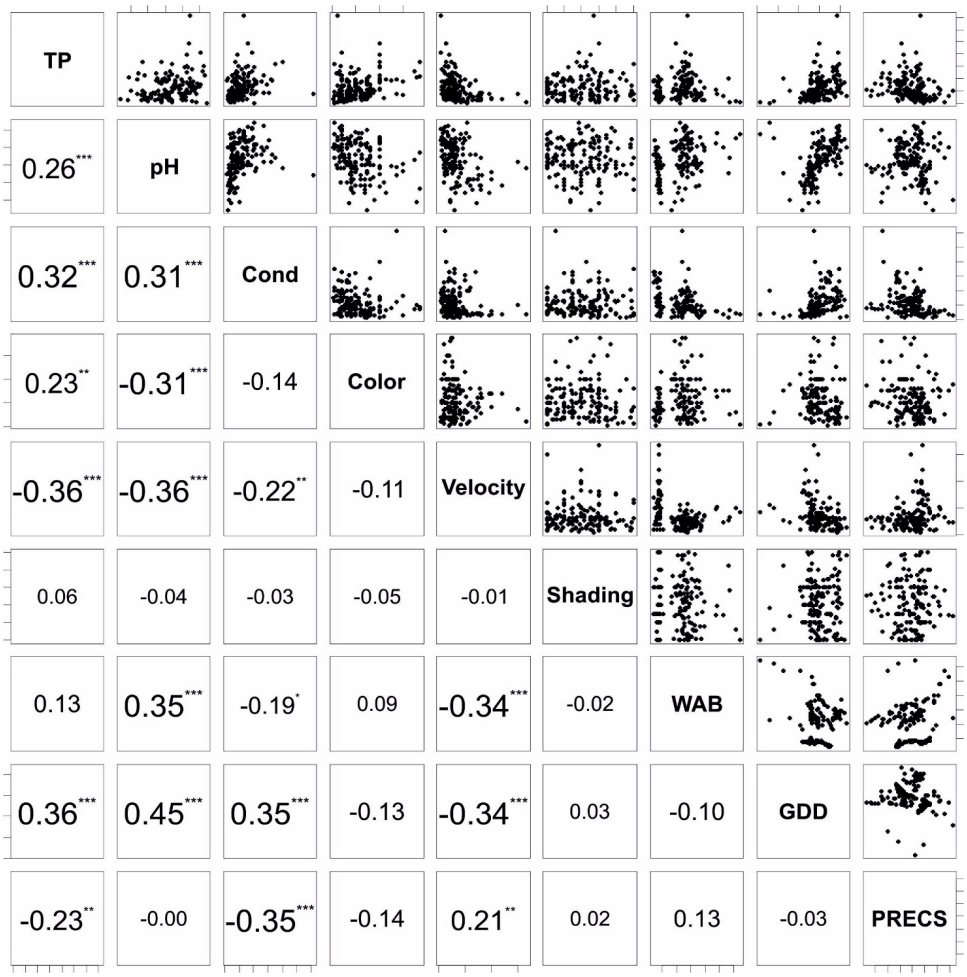


Figure S1 Bivariate matrix for modelled variables from 164 impacted (> 5% anthropogenic land use) stream sites. All pairwise correlations were tested with Spearman's correlation coefficient. The possible covariance between local and climatic variables was low since all correlations were $r_s<|0.50|$. Font sizes are scaled to match the correlation levels. The abbreviations stand for total phosphorus (TP), conductivity (Cond), water balance (WAB), growing degree days (GDD) and summer precipitation (PRECS). Significant codes for *P* values are shown after each correlation: ´***` $P<0.001$, ´**` $P<0.01$, ´*` $P<0.05$, ´` $P\geq0.05$.

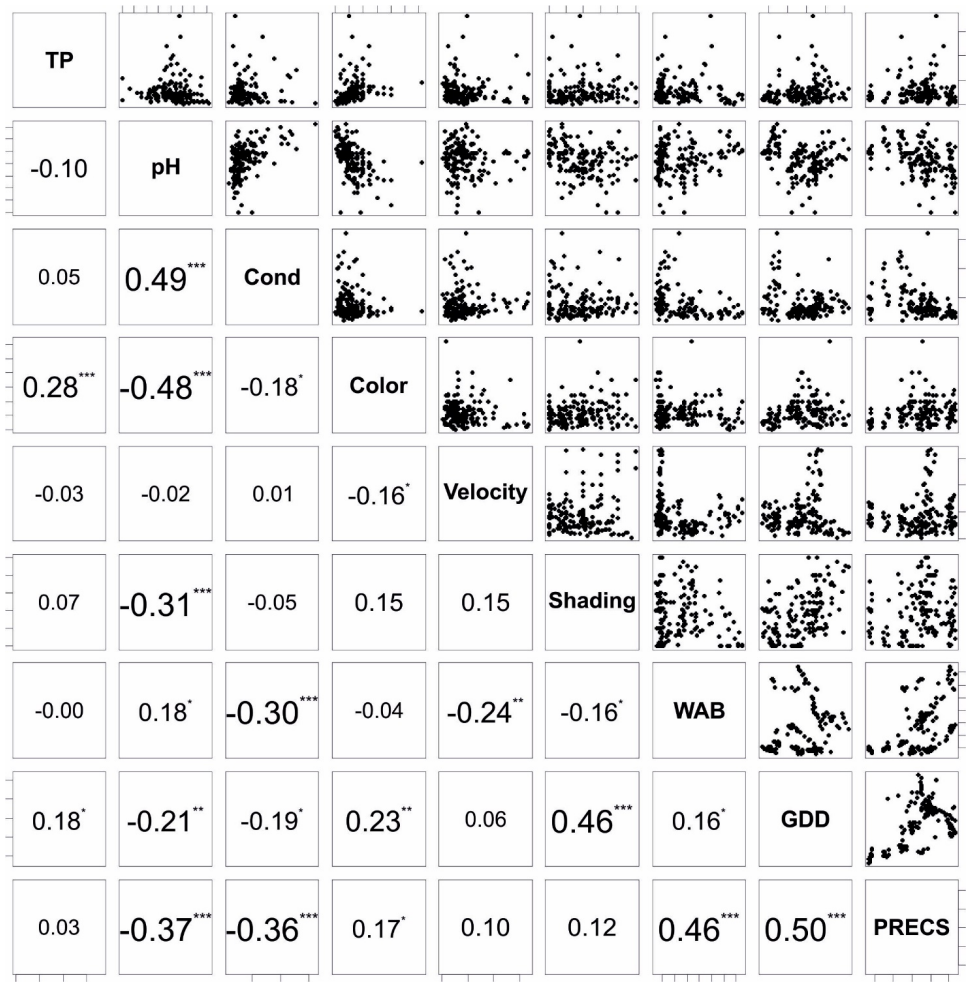


Figure S2 Bivariate matrix for modelled variables from 164 pristine (< 5% anthropogenic land use) stream sites. All pairwise correlations were tested with Spearman's correlation coefficient. The possible covariance between local and climatic variables was low since all correlations were $r_s \leq |0.50|$. Font sizes are scaled to match the correlation levels. The abbreviations stand for total phosphorus (TP), conductivity (Cond), water balance (WAB), growing degree days (GDD) and summer precipitation (PRECS). Significant codes for *P* values are shown after each correlation: ´***` $P<0.001$, ´**` $P<0.01$, ´*` $P<0.05$, `` $P \geq 0.05$.

APPENDIX S2 Additional relationship figures of model performances and predictors.

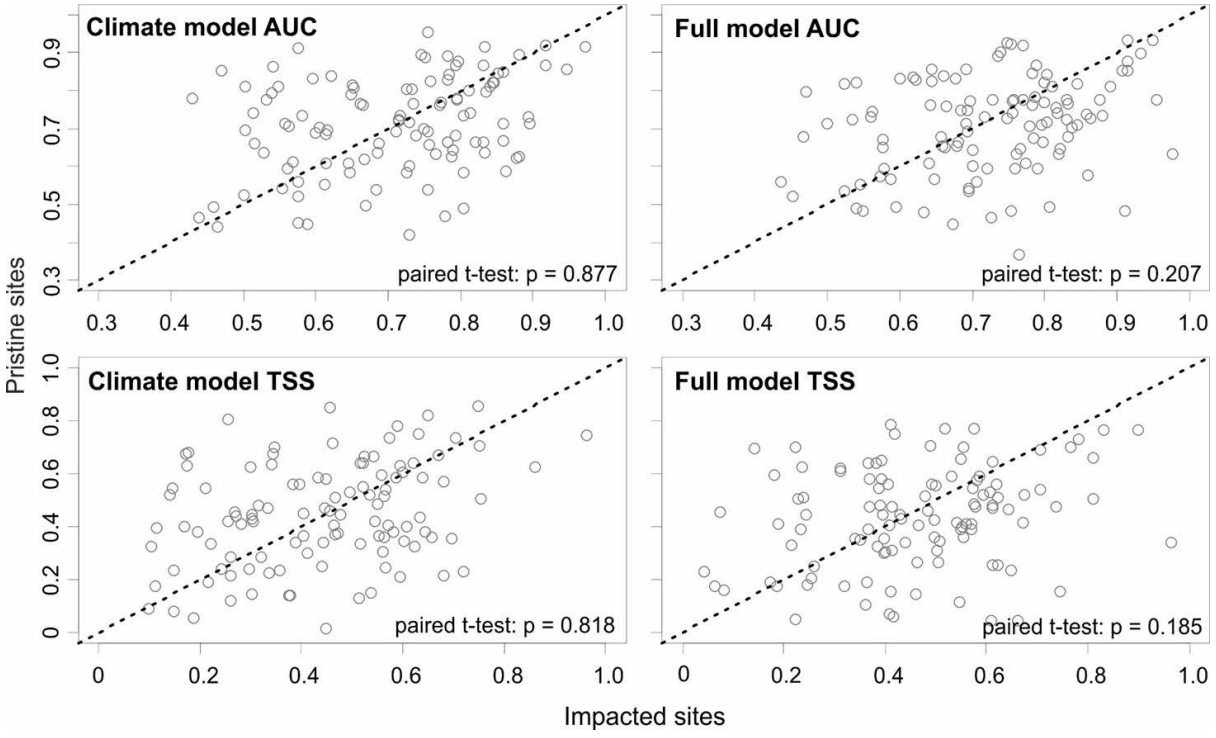


Figure S1 Relationships between the predictive performances of diatom species distribution models in human impacted and pristine sites. Model performances are presented as the area under the curve of a receiver operating characteristic plot (AUC) and the true skill statistic (TSS) values. The models were conducted using boosted regression trees (BRT) as modelling method, and the differences between the site groups are compared in each plot using paired t-test. The climate models consist of three climatic predictors (growing degree days, precipitation and water balance) and full models consist of the three climatic predictors and six additional local environmental predictors: conductivity, total phosphorus, pH, water color, shading by the canopy and current velocity. Dashed lines demonstrate the diagonal line (0, 1).

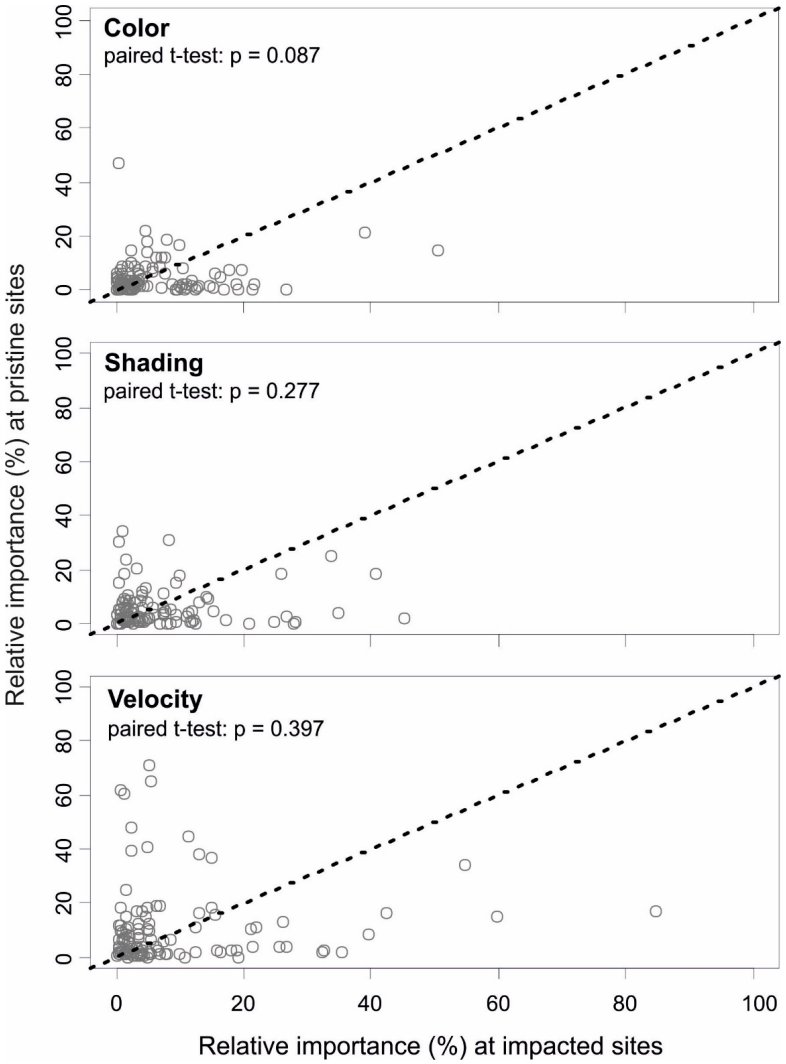


Figure S2 Relationships between the relative importance of three predictors for diatom species distribution in human impacted (> 5% anthropogenic land use) and pristine (< 5% anthropogenic land use) sites. The models were conducted using boosted regression trees (BRT) as modelling method and the full set of predictors. The full models consist of three climatic predictors (growing degree days (GDD), precipitation (PRECS) and water balance (WAB) and six local environmental predictors (conductivity, total phosphorus (TP), pH, water color, shading by the canopy and current velocity). The differences between the site groups are compared in each plot using paired t-test. Solid lines demonstrate the diagonal line (0, 1).

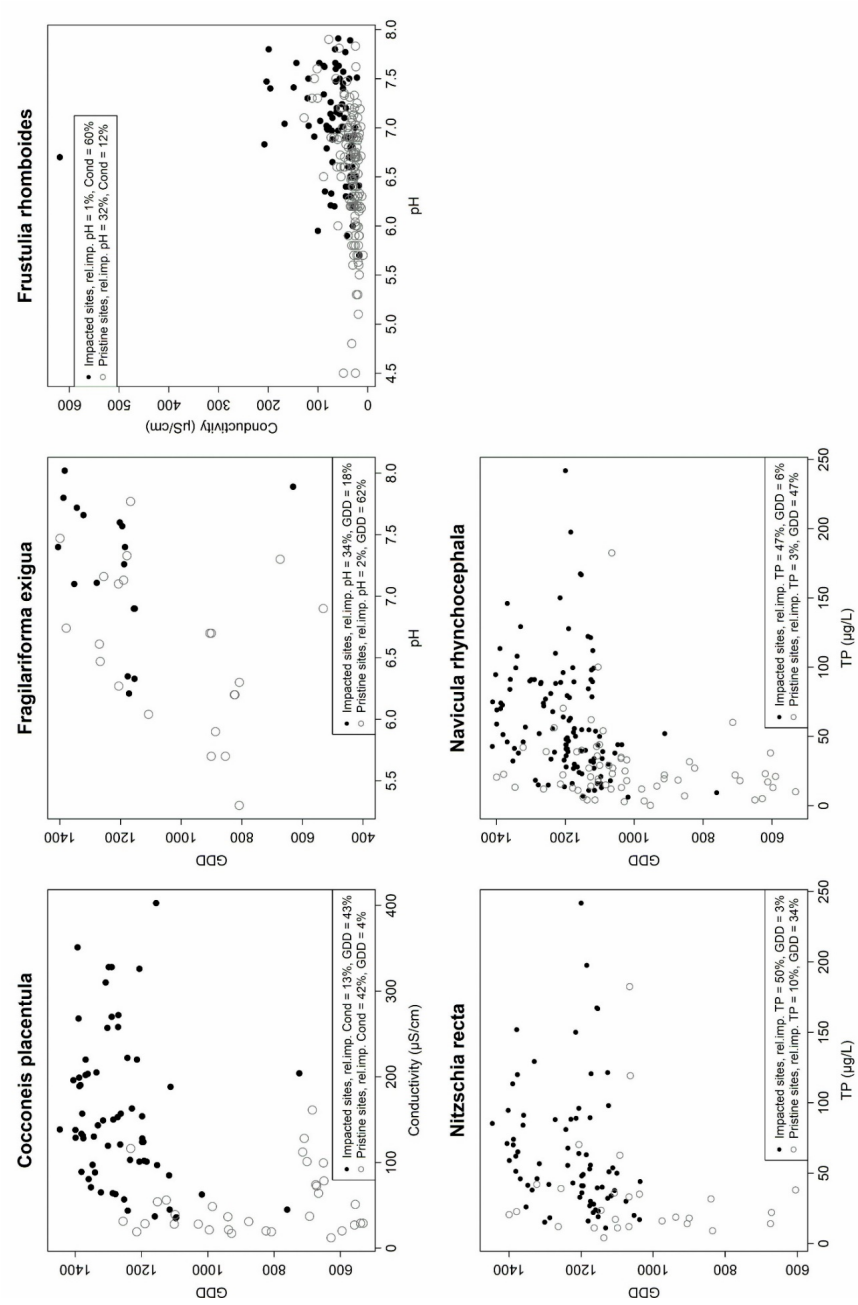


Figure S3 Relationships between two predictor variables at the sites of occurrence of five diatom species in Finnish streams. Occurrences at impacted (> 5% anthropogenic land use) sites are indicated as black dots and occurrences at pristine (< 5% anthropogenic land use) sites as grey circles. The legends show the relative importance of each predictor on diatom species distributions separately at impacted and pristine sites. The abbreviations stand for growing degree days (GDD) and total phosphorus (TP).

Paper III

Pajunen, V., Luoto, M., Soininen, J. 2017. Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities.
Journal of Biogeography 44, 2376-2385.

DOI: 10.1111/jbi.13046

ORIGINAL ARTICLE

WILEY Journal of Biogeography

Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities

Virpi Pajunen | Miska Luoto | Janne Soininen

Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland

Correspondence
Virpi Pajunen, Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland.
Email: virpi.pajunen@helsinki.fi

Funding information
Maj and Tor Nessling foundation

Editor: Jonathan Waters

Abstract

Aim: The diversity and distributions of organisms are jointly influenced by local environment and large-scale variables, such as land cover patterns, dispersal processes and climate. However, the drivers of microbial diversity are complex and the pathways of these drivers' effects are to date largely unresolved, especially in freshwaters. We disentangled the causal direct and indirect effects of climate, land cover and water chemistry on stream diatom communities using hierarchical settings.

Location: Finland, a geographical gradient of *c.* 1200 km (60°–70°N).

Methods: We used structural equation modelling (SEM) to analyse patterns in diatom richness, composition and the uniqueness of species composition in 143 streams. The community composition was defined as the value of the first axis in non-metric multidimensional scaling and the uniqueness of species composition of each site as its local contribution to beta diversity.

Results: Species richness was mainly affected directly by energy and resource availability, increasing with nutrients but decreasing with growing degree days (GDD). The community composition was strongly influenced directly by conductivity and mainly indirectly by anthropogenic land use but also to a lesser degree by GDD (directly and indirectly), indirectly by precipitation and directly by the amount of boreal wetlands. The uniqueness of species composition increased with conductivity but decreased with nutrient concentrations with high number of unique species in southernmost and northernmost sites.

Main conclusions: Our SEM analyses revealed some of the important links between climate, land cover and water chemistry, all of which influenced the microbial diversity. Energy availability has varying indirect and direct effects on diatom communities regardless of local stream conditions, whereas land cover patterns affect aquatic communities most likely indirectly through water chemistry. Collectively, these results suggest that it is important to consider environmental variables simultaneously at different hierarchically ordered scales in macroecological and biogeographical research.

KEYWORDS
beta diversity, catchment, climate, community structure, diatoms, Finland, freshwater, structural equation modelling

1 | INTRODUCTION

The studies of biodiversity patterns seek to understand the mechanisms driving species distributions, interactions between species and the variation in number of species at different spatial scales (Mittelbach, 2012). Depending on the spatial scale, species diversity can be defined as species richness in a local habitat (alpha diversity), the difference in species composition between sites (beta) or regional diversity (gamma) (Whittaker, 1972). Richness and community composition vary not only among local sites but also among biogeographical regions, thus being jointly influenced by local environment as well as large-scale variables, such as climate, evolutionary history and dispersal processes (Heino et al., 2010; Passy, 2010; Potapova & Charles, 2002; Verleyen et al., 2009; Vyverman et al., 2007). The occurrence of species in a certain location depends on hierarchically ordered filtering processes, consisting of climatic factors, geology, physical barriers, historical events, dispersal processes, biotic interactions and the physico-chemical nature of the local habitat, operating at different spatial (i.e. global, regional and local) and temporal scales (Cox, Moore, & Ladle, 2016; Poff, 1997).

In freshwater environments, filters comprise local abiotic (e.g. nutrients, light, water pH and habitat i.e. substrate availability) and biotic (e.g. predation, grazing, competition) variables, variables acting at intermediate scales (e.g. catchment land cover, soil type, geology), and drivers operating at larger spatio-temporal scales (climate, dispersal and historical factors) (Frissell, Liss, Warren, & Hurley, 1986; Stevenson, 1997). According to a hierarchy framework in streams (Frissell et al., 1986), larger scale geomorphic processes constrain the meso- and microscale patterns. Thus, fluvial ecosystems consist of several spatially nested subsystems and are firmly connected with their catchments. The significance of local conditions is obvious especially to passively dispersed organism groups, such as plants and microalgae, which are more related to their local habitat than actively dispersed taxa. Nevertheless, the local conditions are influenced by factors operating at larger scales as e.g. instream conductivity, water pH and nutrient concentrations are strongly related to land use such as agriculture and forestry, and stream flow variability (i.e. disturbance regime) is affected by topography, the degree of urbanization (impervious surfaces) and climate, for example. Land cover has a strong but complex influence on stream abiotic conditions (Allan, 2004). Moreover, land cover may also impact aquatic communities such as bacteria (Lear & Lewis, 2009), lake plankton (Soininen & Luoto, 2012), diatoms (Leland & Porter, 2000; Pan et al., 2004) and macroinvertebrates (Sponseller, Benfield, & Valett, 2001). The effect of land cover on biotic communities is amplified during flooding events as run-off increases nutrient and sediment loadings from land to stream, thus highlighting the significance of seasonality and climate in general on this matter (Jeppesen et al., 2009; Pan et al., 2004).

Whereas the spatial patterns in biodiversity of macro-organisms have been widely studied, the distributional patterns of micro-organisms have only recently gained more attention (Astorga et al., 2012; Martiny et al., 2006; Nemergut et al., 2013). Due to the small body size and correspondingly high rates of reproduction and dispersal, a traditional view of ubiquitous dispersal suggests that microbes have to pass only the local habitat filter in order to occur at a site. Thus, the observed patterns of species' occurrences would reflect merely the influence of local environmental variation (Baas-Becking, 1934; Beijerinck, 1913; Finlay, 2002). However, growing evidence suggests that also large-scale factors such as climate (Berthon, Alric, Rimet, & Perga, 2014; Leira & Sabater, 2005; Pajunen, Luoto, & Soininen, 2016; Weckström, Korhola, & Blom, 1997), history (Vyverman et al., 2007) or dispersal-related processes (Soininen, Paavola, & Muotka, 2004; Verleyen et al., 2009) shape unicellular diatoms, and other microbes (e.g. Papke, Ramsing, Bateson, & Ward, 2003 [cyanobacteria]; Woodcock et al., 2007 [bacteria]; Tedersoo et al., 2014 [fungi]).

Diatoms are typically one of the most diverse and abundant group of algae in freshwaters (Stevenson, Bothwell, & Lowe, 1996). As they respond predictably to many water chemistry variables, they are widely used as indicators of long-term environmental conditions in ecological assessment (Smol & Stoermer, 2010; and references therein). They also maintain vital ecosystem processes (e.g. nutrient cycling) and comprise an important part of food webs in streams (Allan & Castillo, 2007; and references therein). Even if the factors affecting freshwater diatoms at different spatial scales have been investigated before, the studies typically analyse only the direct and joint effects of environmental and spatial variables on diatoms using e.g. direct ordination with variation partitioning, whereas causality and indirect effects remain unresolved (but see Passy, 2010). New statistical methods allow the quantification of direct impacts of environmental factors together with indirect effects on biota, and test the causality of these effects. Structural equation modelling (SEM) is a highly useful tool for exploring the causality of the direct and indirect effects of factors operating at multiple scales. Unlike many more traditional modelling approaches, which demonstrate mere correlation between variables, SEM is a technique that links a set of predictors and response variables in a causal network (Lefcheck, 2016). The paths in this network represent hypothesized causal relationships between variables and the variables can perform simultaneously as both predictors and responses, thus acting as mediators between explanatory and response variables.

The aim of this study is to use SEM to analyse patterns in stream diatom richness, composition and the uniqueness of species composition using hierarchical settings of stream sites within catchments. We hereby address the following questions using a diatom data set of 143 stream sites in Finland: (1) What are the direct effects of local and climatic factors on stream diatom communities? (2) Do climate and land cover variables have indirect effects on diatoms through local variables? We characterize diatom communities using species richness, community composition and the uniqueness of species composition at sites, measured as local contribution to beta diversity (Legendre & De Cáceres, 2013). The answers would give novel insights into the complex pathways of the variables affecting microbial community composition and diversity in boreal streams.

2 | MATERIALS AND METHODS

2.1 | Data sampling and analysis

A data set of 143 stream sites was collected from streams with independent catchments (one site per stream) in Finland (60°–70°N, 20°–32°E) between 1986 and 2001 covering all the five ecoregions of Finland but also long gradients of water quality (Soininen et al., 2004), land cover and climatic variables (Pajunen et al., 2016) (see Appendix S1 in Supporting Information, Table S1.1). Sites were sampled once and all sampling was conducted during the low flow conditions between June and August using identical sampling methods. Each study site was divided into ten transects perpendicular to the flow with an even coverage of the whole study section. Current velocity, shading by the riparian canopy and stream width were measured from each transect. Five to 10 pebble-to-cobble (5–15 cm) sized stones were collected from the study area and diatoms were sampled by brushing the stones with a toothbrush, according to the recommendations of Kelly et al. (1998). The diatom suspension was pooled into a small plastic bottle and preserved in ethanol (70%). At most of the sites, water samples were taken simultaneously with the diatom samples and analysed for total phosphorus (TP), pH, conductivity and water colour. For less than 20% of the sites, water chemistry data were taken from the national water quality database, using results from the nearest sampling occasion.

At the laboratory, diatom samples were cleaned from organic material using wet combustion with acid (HNO_3:H_2SO_4; 2:1) and mounted in Naphrax or Dirax. A total of 250–500 valves per sample were identified and counted using phase contrast light microscopy (magnification 1,000×) (Figure 1). Species were identified according to Krammer and Lange-Bertalot (1986–1991) and Lange-Bertalot and Metzeltin (1996) by two analysts who harmonized species identification.

2.2 | Catchments and land cover

For each stream site, upstream catchment area was defined by calculating the patterns of flow direction and accumulation from digital elevation model (grid resolution 10 × 10 m, National Land Survey of Finland, 2013) in ArcGIS 10.3.1 software. The catchment size, the slope variability and land cover were further calculated for each catchment. Catchment classifications of land use and cover were obtained from CORINE Land Cover 2000 data (25 × 25 m, Finnish Environment Institute 2005). Base maps from the catchment areas of the sites collected during 1986 were compared to CORINE land cover data and no significant changes in land use were observed between 1986 and 2000. CORINE land cover classes were merged into nine categories best describing the study area: artificial areas, agricultural areas, broad-leaved forest, coniferous forest, mixed forest, shrubs, open spaces with little or no vegetation, wetlands and water bodies. Artificial and agricultural areas were further combined to create an anthropogenic land use class. The data set was divided into two subsets by the amount of anthropogenic land use in the catchment area: reference sites (<5%, n = 97) and impacted sites (≥5%, n = 46).

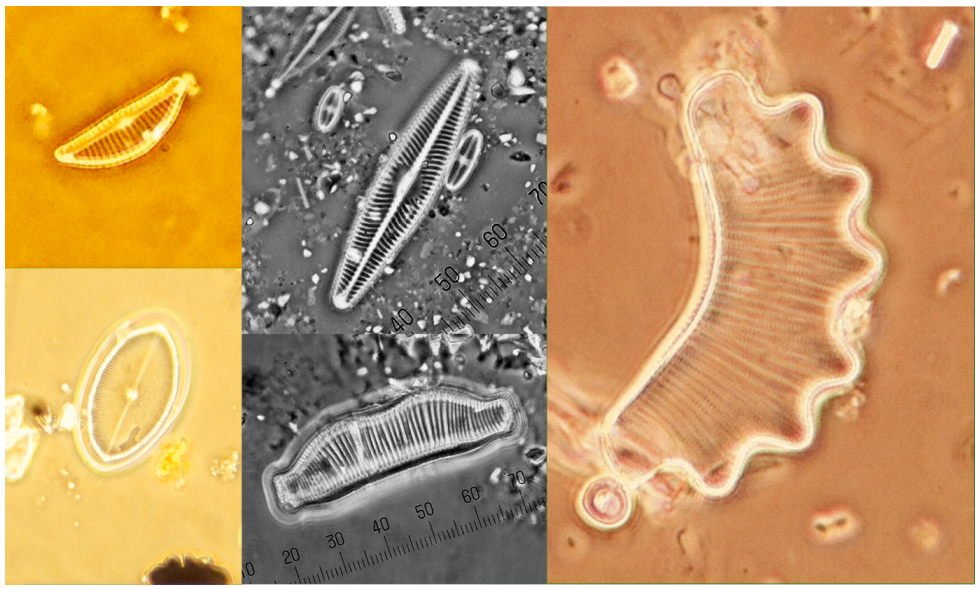


FIGURE 1 Some diatom species in Finnish streams. The scale is in μm. Photo credit: Virpi Pajunen

2.3 | Climatic variables

The climatic data covering averages from years 1981–2010 were obtained from Finnish Meteorological Institute. Climate data were related to latitude, longitude and elevation of each study site using multiple linear regression and downscaling climate data from 10 × 10 km resolution grid to the study site (Finnish Meteorological Institute; Venäläinen & Heikinheimo, 2002). We chose to use growing degree days (GDD, adjusted to 5°C) and summer precipitation sum from May to September (PRECS) as climatic variables in this study as they are known to have a strong impact on diatom assemblages (Pajunen et al., 2016). GDD (i.e. the accumulation of heat units) represents both the length and the mean air temperature of the growing season and it has a strong negative correlation with latitude in our data set ($r_p = -.96$, $p = .000$, see Appendix S1, Fig. S1.1). PRECS represent the moisture availability, which in lotic environments refers to the amount of recharge and run-off. GDD and precipitation may represent also the overall productivity of stream environment including its catchment as they are found to be strong predictors of NDVI (Li, Tao, & Dawson, 2002; Wang, Price, & Rich, 2001).

2.4 | Statistical analysis

All the statistical analyses were performed in R environment (version 3.3.2; R Development Core Team, 2008). Species richness, community composition and uniqueness were selected as response variables in SEMs. Species richness was defined as the number of taxa at each site, and the community composition as the value of the first axis in non-metric multidimensional scaling (NMDS). The first axis represents the most important variation in the species data and its values are thus robust in describing the community composition. Here, the NMDS1-variable represents the difference in community composition between sites as sites with values in opposite ends of the axis (the most negative and positive ends) had widely different species composition. NMDS was performed using relative abundance data in "vegan" package (Oksanen et al., 2015) with three axes (stress value = 0.18). The uniqueness of species composition of each site was defined as its local contribution to beta diversity (LCBD, Legendre & De Cáceres, 2013). By using this measure, we calculate the contribution of the sites to the overall beta diversity of the data set. LCBD was derived from a total variation (beta diversity) in a species-by-site community matrix based on Hellinger-transformed data using the "beta.div" function in R. Significance of the LCBD values was then tested for using 999 permutations. All procedures followed the methods developed by Legendre and De Cáceres (2013).

Data were then examined by performing a principal component analysis for all standardized environmental variables. A redundancy analysis was performed for species assemblage data and generalized linear models were ran to examine the most important environmental variables for species richness, NMDS1 and LCBD (see Appendix S2, Figs. S2.4-5, Tables S2.2-4). Based on these results, we chose the most important environmental variables operating at three different spatial scales for the SEMs: local scale (TP and conductivity), intermediate scale (anthropogenic land use and wetlands) and large-scale (GDD and PRECS) variables as predictors. We adjusted the number of predictors (two from each spatial level) to the sample size to avoid model complexity which would compromise the reliability and hinder the interpretation of the models (Lefcheck, 2016). We included conductivity in the models instead of water pH as it is more conservative and has been identified as a stronger explanatory variable for diatom community composition than pH in boreal streams (Soininen et al., 2004). Nevertheless, we conducted alternative SEMs including pH as a seventh environmental variable to account for the influence of pH on diatom communities (see Appendix S3, Tables S3.5-6).

We used SEMs to investigate the causal links between the independent factors and dependent variables. SEMs are probabilistic models that allow the simultaneous investigation of multiple causal pathways within a single network (Lefcheck, 2016). Data were checked for nonlinear relationships and strong covariances (see Appendix S1, Figs. S1.2-3), and normalized and standardized prior analysis. The models were run using the "piecewiseSEM" package in R developed by Lefcheck (2016). In piecewise SEM, paths are structured as a set of separate linear equations, which are evaluated individually using local estimation. This allows more freedom in sample sizes and sampling designs compared to other SEM methods where equations are solved simultaneously (Shipley, 2000, 2009). For example, piecewise SEM allows smaller sampling sizes and non-normal distribution of variables. Despite these differences, the interpretation of piecewise SEM is similar to other SEMs (Lefcheck, 2016).

Models were built separately for each of the three dependent variables: richness, NMDS1 scores and LCBD. For each model, a basic model was built including GDD and PRECS as exogenic variables and anthropogenic land use (anthro), wetlands, TP and conductivity as endogenic variables. All potential causal links between variables, as well as the quadratic terms of GDD and PRECS, were included in the original models. Different model structures were then tested by gradually removing non-significant pathways maintaining the causal structure of the models. To account for significant nonlinear relationships, composite variables were made using the coefficient estimates of linear and nonlinear terms of GDD and PRECS by multiplying the terms by their estimates and then adding them as aggregate. Composites were made for the following paths: GDD—anthro, GDD—wetlands, GDD—NMDS1 and for PRECS—conductivity.

Alternative SEMs were conducted including geographical coordinates as additional variables (see Appendix S3, Tables S3.5-6). There were eight variables in models with geographical coordinates. The latter models comprised four levels of predictive variables: spatial, climatic, land cover and local variables.

Models were rejected or accepted by the criterion of parsimony and goodness-of-fit using Fisher's C, where $p > .05$ represents a good fit. Spatial autocorrelation was taken into account in all models except models with geographical coordinates by using "spatialCorrect" function, which corrects the standard error for all endogenous variables using Moran's *I*.

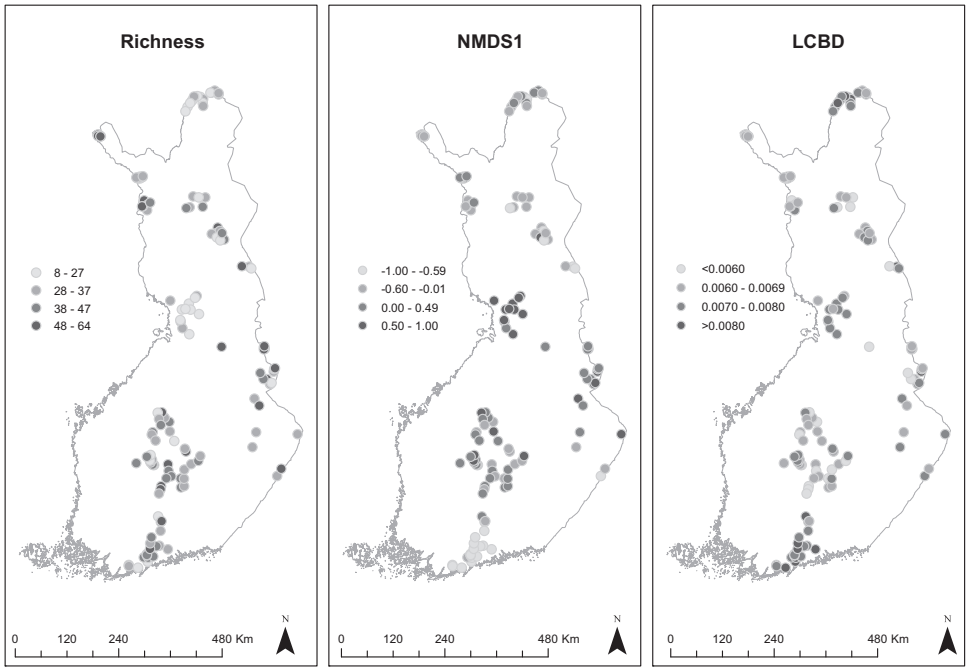


FIGURE 2 The sampling sites and the distribution of diatom species richness, community composition (represented as the scores of the first axis on non-metric multidimensional scaling; NMDS1) and uniqueness of species composition (represented as values of local contribution to beta diversity; LCBD) in the study area located in Finland, Northern Europe (60°–70°N)

3 | RESULTS

The most species-rich sites were mainly randomly distributed, yet some richness hotspots were located in the northern and eastern Finland (Figure 2). The regions with species-poor sites were mainly located in the northern and central Finland, typically in areas characterized by large wetlands or strong human impact.

The sites that differed the most in their community composition from each other were situated mainly in the coastal area in the southern Finland (the negative end of the NMDS1-axis) and in the central Finland (the positive end of the axis) (Figure 2). The sites in the negative end of the NMDS1-axis (dominated by species such as *Surirella brebissonii* [Krammer & Lange-Bertalot], *Navicula cryptotenella* [Lange-Bertalot] and *Nitzschia palea* [Kützing] W. Smith) consisted mainly of impacted sites (anthropogenic land use ≥5%) having also high values of GDD and conductivity (Figure 3) indicating eutrophic conditions. However, the sites in the positive end were mostly reference sites (anthropogenic land use <5%) with low water pH and humic waters (dominated by species such as *Eunotia bilunaris* ([Ehrenberg] Schaarschmidt), *Eunotia incisa* (W. Smith ex W. Gregory) and *Frustulia rhomboides* ([Ehrenberg] De Toni), having the highest relative amount of wetlands in their catchment (Figure 3).

The sites with highest LCBD were located in the southern coastal area, the northernmost Finland and in few sites in the eastern Finland (Figure 2). Species richness and LCBD had a negative correlation in the data ($r_p = -.53$, $p = .000$, see Appendix S1, Figure S1.1).

For species richness, the most parsimonious SEM included direct effects of GDD (negative effect), PRECS (positive), wetlands (negative) and TP (positive) (Figure 4a) with GDD having the strongest effect. However, GDD had an indirect positive influence on richness via TP and catchment level variables.

For NMDS1 scores, the best SEM included the direct effects of GDD (unimodal), anthropogenic land use (negative), wetlands (positive) and conductivity (negative) (Figure 4b). Conductivity had by far the strongest effect. Unimodal effect with GDD indicated that the communities with average value of GDD differed most from the communities with highest and lowest GDD. TP did not have a significant effect on NMDS.

For LCBD, the best SEM included the direct effects of conductivity, having a strong and positive effect which saturated at higher levels, and TP, which had a weaker negative effect (Figure 4c). The climatic and catchment scale variables had only indirect effects on LCBD.

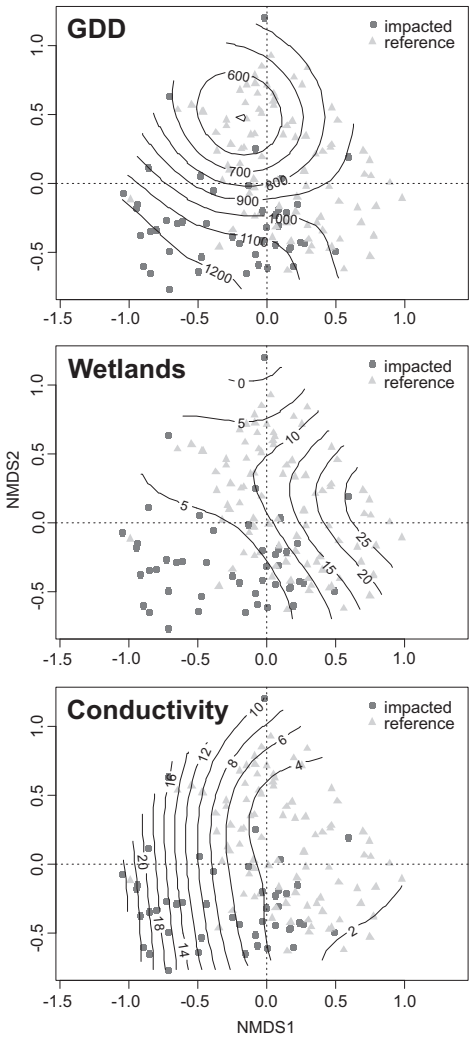


FIGURE 3 The position of the values of growing degree days (GDD), wetlands and conductivity on the non-metric multidimensional scaling (NMDS) axes shown as contour lines. NMDS is based on diatom abundance data in Finland. The sampling sites are divided into impacted sites (anthropogenic land use ≥5%, shown as dark grey dots) and reference sites (anthropogenic land use <5%, shown as light grey triangles)

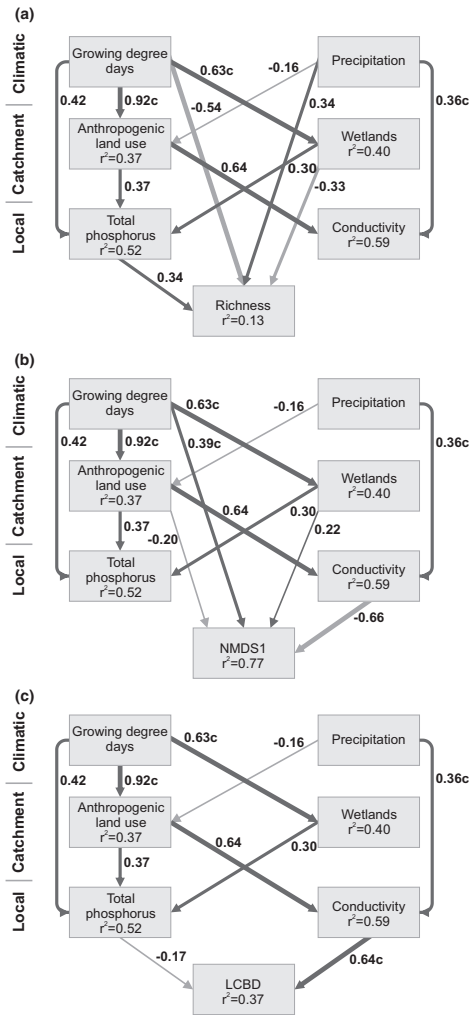


FIGURE 4 Structural equation models of the effects of climatic, land cover and water chemistry variables on diatom communities in Finnish streams. Separate models represent the influences on (a) species richness, (b) community composition (measured as the scores of the first axis on non-metric multidimensional scaling; NMDS1), and (c) the uniqueness of species composition (measured as values of local contribution to beta diversity; LCBD). Only significant standardized path coefficients are shown ($p < .05$). The width of the arrows represents strength of influence between variables

The SEMs were best at explaining the variation in community compositions ($r^2 = .77$ for NMDS1), while the models for LCBD and species richness had lower coefficient of determination ($r^2 = .37$ for LCBD and $r^2 = 0.13$ for richness).

Adding pH in the alternative SEMs did not increase the amount of explained variation in biotic response variables in any of the models (see Appendix S3). Water pH did not have a statistically

significant effect on richness and its effects on NMDS1 and LCBD were only minor. Geographical coordinates, however, increased the amount of explained variation in richness (r^2 = .20) and NMDS1 (r^2 = .78) models but decreased it in LCBD models (r^2 = .18). Latitude had a strong direct effect on richness and NMDS1, and longitude on NMDS1. However, due to the strong correlations between geographical coordinates and other variables (e.g. latitude—GDD, r = −.96, see Appendix S1, Figure S1.1), model robustness decreases.

4 | DISCUSSION

Diatom species richness and community composition were related with local, catchment level and climatic factors showing causality through direct effects but also showed multiple indirect effects. On the uniqueness of species composition, only local variables had direct effects, but uniqueness was affected indirectly by climate and catchment level variables. The models showed direct effects of precipitation and land cover on species richness and community composition, indicating that large-scale variables tend to reflect the aquatic conditions at longer time-scales than snapshot measures of local variables, or large-scale effects are mediated through some unmeasured local environmental variable (Soininen & Luoto, 2012), and thus, have a direct effect on communities. Our results indicate that climatic and land cover variables may act as valuable proxies explaining the variation on diatom species richness and community composition. Nevertheless, highlighting the importance of climate, GDD can have a direct influence on stream diatoms via water temperature, which typically correlates with air temperature (Mohseni & Stefan, 1999). This demonstrates not only the importance of indirect effects of climate and land cover on stream biota, but also the fact that large-scale variables may affect diversity and distributional patterns of microbial stream biota also directly. This finding has important implications also for practical biomonitoring purposes and stream conservation that currently largely neglect large-scale variables (Heino, 2013).

For species richness, perhaps the most interesting finding was that GDD had a strong negative direct effect on diatom richness. This finding was surprising, as regions with higher GDD (energy availability) typically have higher primary production and species richness (Hawkins et al., 2003) and also nutrient availability (Rouse et al., 1997). The found negative effect of GDD on diatom richness may reflect the fact that in colder climates, water temperatures are lower, and consequently, diatom richness higher. This finding is supported by previous experimental work, where diatom diversity was higher at lower stream temperatures (Gudmundsdottir, Olafsson, Palsson, Gislason, & Moss, 2011). The underlying reason could be that high stream temperatures can decrease some diatom populations perhaps due to interspecific competition (Piggott, Salis, Lear, Townsend, & Matthaei, 2015) thus resulting in lower diversity through local extinctions. Also, the presence of priority effects (i.e. the early colonizers pre-empt niche space) can result in resistant communities comprising highly abundant species that can inhibit later colonization of other species (Andersson, Berga, Lindström, & Langenheder, 2014; Louette & De Meester, 2007). In addition, Soininen, Jamoneau, Rosebery, and Passy (2016) recently found that there was a weak latitudinal increase in stream diatom richness globally suggesting a negative relationship between richness and temperature, whereas Passy (2010) found a distinct U-shaped latitudinal pattern in stream diatom richness. Collectively, these findings suggest that, at the very least, there may not be a simple positive relationship between richness and energy availability in diatoms that have been found in many other organism groups (Hillebrand, 2004).

The global and continental diatom richness patterns documented by Soininen et al. (2016) and Passy (2010) were possibly linked to corresponding patterns in catchment and stream characteristics (e.g. distribution of wetlands and forests, soil composition) influencing resource supply (i.e. DOC and iron concentrations increased with the amount of temperate wetlands) (Passy, 2010) and water pH (Soininen et al., 2016). However, in our data set, wetlands comprised mainly boreal peatlands resulting in relatively high DOC concentrations (indicated by brown water colour) and low pH and conductivity of the stream water. Although wetlands had a positive effect on diatom richness through elevated TP concentrations in present data, the direct and overall effect of wetlands remained negative disagreeing thus with the conclusions of Passy (2010). Such a disagreement probably indicates the differences between temperate and boreal wetlands as the first may increase nutrient supply while the latter typically decreases water pH, light availability and conductivity. Furthermore, in naturally brown water streams rich with humic compounds, iron that is important nutrient for diatoms can be bound with chelated humic acids (Collier, Ball, Graesser, Main, & Winterbourn, 1990; Winterbourn & Collier, 1987) and thus its bioavailability is uncertain.

Diatom community composition was strongly affected by GDD and conductivity associated with anthropogenic land use (the negative end of the NMDS axis 1), and on the other hand, by the amount of wetlands (the positive end) although to a much lesser degree. The strong effect of anthropogenic land use on conductivity indicates that, in present data, high conductivity is typically related to intensive human impacts in the catchments. The unimodal direct effect of GDD on community composition peaked at mean GDD values, representing the region in the central Finland with plains, seasonal floods and abundant peatlands, consistent with the effect of wetlands on diatom communities. The linkage between climate and peatlands is evident as ecosystem functioning and the formation of boreal peatlands rely on the critical connections between temperature, precipitation and permafrost (Lavoie, Paré, & Bergeron, 2005).

For the uniqueness of species composition at sites, it was evident that both the southernmost and northernmost sited contributed most to the overall beta diversity in the data and that such pattern was clearly related to conductivity having positive effect on LCBD. This finding agrees with Tonkin, Heino, Sundermann, Haase, and Jähnig (2016) who found that LCBD for stream macroinvertebrates was mostly explained by local habitat conditions. Intuitively, large

climatic and environmental variation would lead to more unique sites near the ends of the gradients, thus overall increasing the number of unique species. However, in our study, climatic factors or geographical coordinates had only indirect effects on LCBD indicating that their effects were mediated by local variables. In our data, high conductivity from anthropogenic sources indicates the presence of pollutants and stream degradation, thus favouring species tolerant to such conditions. This contributes to total beta diversity by including unique species not found elsewhere in the study area. On the contrary, TP, increasing with energy availability (GDD) had a slight negative effect on the community uniqueness, indicating that slightly more of the unique species were found in nutrient-poor streams than in eutrophic streams, which tend to harbour more typical species in the data. Such unique species tolerate well cold temperatures and overall harsh and nutrient-poor conditions in higher latitudes with low productivity and biomasses (Wang, Soininen, He, & Shen, 2012). Hence, we conclude that either strong anthropogenic impacts or lack of available resources has selected specialist species that have high contribution to beta diversity in our data set. Conversely, at the intermediate productivity, generalists may out-compete these more unique species.

The negative correlation between LCBD and species richness indicated that high uniqueness of species composition was often related to low number of species, highlighting that sites with high LCBD were occupied with specialized species tolerant of harsh conditions. In fact, according to Legendre and De Cáceres (2013), large LCBD values together with low species richness may reveal sites that have unusual species combinations having thus high conservation value, or on the contrary, it may indicate sites that have degraded and need restoration. In our study area, both of these may occur as uniqueness was highest in species-poor high conductivity sites with strong human impact but also in harsh low-nutrient pristine sites in remote regions. However, the occurrence of priority effects cannot be entirely ruled out as species-poor sites are typically dominated by one or more abundant species.

When comparing the explanatory power of SEM models, community composition was best explained (r^2 = .77) indicating the great importance of included variables such as conductivity (Potapova & Charles, 2003; Soininen et al., 2004), climate (Pajunen et al., 2016) and human impact (Leland & Porter, 2000; Pan et al., 2004) on diatom communities. Moreover, our analyses were conducted with abundance data, which may increase the explanatory power compared to presence–absence data (Heino et al., 2010). Uniqueness of species composition was explained only moderately (r^2 = .37), which implies that some of the important factors governing beta diversity were missing from our models. Finally, the explained variation of species richness in SEM was the smallest (r^2 = .13) suggesting that microbial richness patterns are complex, driven by great number of variables and are typically relatively stochastic in time and space due to the small body size of microbes and their vulnerability to disturbances (Farjalla et al., 2012; Soininen, Korhonen, & Luoto, 2013). Additionally, the richness estimates are also affected by sampling effect as the sampling of relatively limited area in the field and counting of mere 500 valves per sample may not detect all species occurring at site, especially at sites with high diatom species richness (Heino & Soininen, 2005).

In conclusion, we demonstrated using SEM models that stream diatoms are affected by the wide range of environmental factors operating at different spatial scales. The effect of climatic factors and land cover can be strong even if the influence on biota may be largely indirect and reflected through local factors, i.e. water chemistry. We emphasize, however, that climate may have also clear direct effect on diatoms, regardless of local factors. Diatom species richness was mainly governed by GDD and productivity increasing with available nutrients but decreasing with temperature. Richness was also low in extreme conditions such as in brown water or under high human impact. The community composition on the other hand, was affected by conductivity and human impact but also by the unique stream conditions in the central Finland. The high uniqueness of species composition was mainly found in gradient ends of the study area with high number of unique species contributing most to beta diversity at southernmost and northernmost study sites. The present study revealed that not only local variables have direct effects on diatoms but rather that large-scale variables may also have direct causal influence on diatom communities regardless of local factors. This is an important finding, especially due to the projected climate change, to consider not only in basic research of macroecology and biogeography but also in freshwater bioassessment and conservation programs that are typically guided only by local variables.

ACKNOWLEDGEMENTS

This project was funded by Maj and Tor Nessling foundation.

REFERENCES

Allan, J. D. (2004). Landscapes and riverscapes: The influence of land use on stream ecosystems. *Annual Review of Ecology and Systematics*, *35*, 257–284.

Allan, J. D., & Castillo, M. M. (2007). *Stream ecology: Structure and function on running waters* (2nd ed.). Dordrecht: Springer.

Andersson, M. G. I., Berga, M., Lindström, E. S., & Langenheder, S. (2014). The spatial structure of bacterial communities is influenced by historical environmental conditions. *Ecology*, *95*, 1134–1140.

Astorga, A., Oksanen, J., Luoto, M., Soininen, J., Virtanen, R., & Muotka, T. (2012). Distance decay of similarity in freshwater communities: Do macro- and microorganisms follow the same rules? *Global Ecology and Biogeography*, *21*, 365–375.

Baas-Becking, L. G. M. (1934). *Geobiologie of Inleiding Tot de Milieukunde*. The Hague: Van Stockkum & Zoon.

Beijerinck, M. W. (1913) De infusies en de ontdekking der backteriën. In *Jaarboek van de Koninklijke Akademie van Wetenschappen*. pp 1-28. Amsterdam: Müller.

Berthon, V., Alric, B., Rimet, F., & Perga, M.-E. (2014). Sensitivity and responses of diatoms to climate warming in lakes heavily influenced by humans. *Freshwater Biology*, *59*, 1755–1767.

Collier, K. J., Ball, O. J., Graesser, A. K., Main, M. R., & Winterbourn, M. J. (1990). Do organic and anthropogenic acidity have similar effects on aquatic fauna? *Oikos*, *59*, 33–38.

Cox, C. B., Moore, P. D., & Ladle, R. J. (2016). *Biogeography: An ecological and evolutionary approach*. Chichester: John Wiley & Sons Ltd.

Farjalla, V. F., Srivastava, D. S., Marino, N. A. C., Azevedo, F. D., Dib, V., Lopes, P. M., … Esteves, F. A. (2012). Ecological determinism increases with organism size. *Ecology*, *93*, 1752–1759.

Finlay, B. (2002). Global dispersal of free-living microbial eukaryote species. *Science*, *296*, 1061–1063.

Frissell, C. A., Liss, W. J., Warren, C. E., & Hurley, M. D. (1986). A hierarchical framework for stream habitat classification: Viewing streams in a watershed context. *Environmental Management*, *10*, 199–214.

Gudmundsdottir, R., Olafsson, J. S., Palsson, S., Gislason, G. M., & Moss, B. (2011). How will increased temperature and nutrient enrichment affect primary producers in sub-Arctic streams? *Freshwater Biology*, *56*, 2045–2058.

Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J.-F., Kaufman, D. M., … Turner, J. R. G. (2003). Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, *84*, 3105–3117.

Heino, J. (2013). The importance of metacommunity ecology for environmental assessment research in the freshwater realm. *Biological Reviews*, *88*, 166–178.

Heino, J., Bini, L. M., Karjalainen, S. M., Mykrä, H., Soininen, J., Vieira, L. C. G., & Diniz-Filho, J. A. F. (2010). Geographical patterns on microorganismal community structure: Are diatoms ubiquitously distributed across boreal streams? *Oikos*, *119*, 129–137.

Heino, J., & Soininen, J. (2005). Assembly rules and community models for unicellular organisms: Patterns in diatoms of boreal streams. *Freshwater Biology*, *50*, 567–577.

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *The American Naturalist*, *163*, 192–211.

Jeppesen, E., Kronvang, B., Meerhoff, M., Søndergaard, M., Hansen, K. M., Andersen, H. E., … Olesen, J. E. (2009). Climate change effects on runoff, catchment phosphorus loading and lake ecological state, and potential adaptations. *Journal of Environmental Quality*, *38*, 1930–1941.

Kelly, M., Cazaubon, A., Coring, E., Dell'Uomo, A., Ector, L., Goldsmith, B., … Vizinet, J. (1998). Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology*, *10*, 215–224.

Krammer, K., & Lange-Bertalot, H. (1986–1991) *Bacillariophyceae. Süßwasserflora von Mitteleuropa 2 (1-4)*. Stuttgart: Gustav Fischer Verlag.

Lange-Bertalot, H., & Metzeltin, D. (1996) *Iconographica diatomologica, Volume 2. Indicators of oligotrophy. 800 taxa representative of three ecologically distinct lake types: Carbonate buffered, oligodystrophic, weakly buffered soft water*. Koenigstein: Koeltz Scientific Books.

Lavoie, M., Paré, D., & Bergeron, Y. (2005). Impact of global change and forest management on carbon sequestration in northern forested peatlands. *Environmental Reviews*, *13*, 199–240.

Lear, G., & Lewis, G. D. (2009). Impact of catchment land use on bacterial communities within stream biofilms. *Ecological Indicators*, *9*, 848–855.

Lefcheck, J. S. (2016). PiecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*, 573–579.

Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters*, *16*, 951–963.

Leira, M., & Sabater, S. (2005). Diatom assemblages distribution in catalan rivers, NE Spain, in relation to chemical and physiographical factors. *Water Research*, *39*, 73–82.

Leland, H. V., & Porter, S. D. (2000). Distribution of benthic algae in the upper Illinois River basin in relation to geology and land use. *Freshwater Biology*, *44*, 279–301.

Li, B., Tao, S., & Dawson, R. W. (2002). Relations between AVHRR NDVI and ecoclimatic parameters in China. *International Journal of Remote Sensing*, *23*, 989–999.

Louette, G., & De Meester, L. (2007). Predation and priority effects in experimental zooplankton communities. *Oikos*, *116*, 419–426.

Martiny, J. B. H., Bohannan, J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., … Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, *4*, 102–112.

Mittelbach, G. G. (2012). *Community ecology*. Sunderland: Sinauer Associates Inc.

Mohseni, O., & Stefan, H. G. (1999). Stream temperature/air temperature relationship: A physical interpretation. *Journal of Hydrology*, *218*, 128–141.

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Billinski, T. M., Stanish, L. F., … Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews*, *77*, 342–356.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., … Wagner, H. (2015). Vegan: Community ecology package. Retrieved from http://cran.r-project.org/web/packages/vegan/index.html (accessed September 2016).

Pajunen, V., Luoto, M., & Soininen, J. (2016). Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography*, *25*, 198–206.

Pan, Y., Herlihy, A., Kaufmann, P., Wigington, J., van Sickle, J., & Moser, T. (2004). Linkages among land-use, water quality, physical habitat conditions and lotic diatom assemblages: A multi-spatial scale assessment. *Hydrobiologia*, *515*, 59–73.

Papke, R. T., Ramsing, N. B., Bateson, M. M., & Ward, D. M. (2003). Geographical isolation in hot spring cyanobacteria. *Environmental Microbiology*, *5*, 650–659.

Passy, S. I. (2010). A distinct latitudinal gradient of diatom diversity is linked to resource supply. *Ecology*, *91*, 36–41.

Piggott, J. J., Salis, R. K., Lear, G., Townsend, C. R., & Matthaei, C. D. (2015). Climate warming and agricultural stressors interact to determine stream periphyton community composition. *Global Change Biology*, *21*, 206–222.

Poff, N. L. (1997). Landscape filters and species traits: Towards mechanistic understanding and prediction in stream ecology. *Journal of North American Benthological Society*, *16*, 391–409.

Potapova, M., & Charles, D. F. (2003). Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. *Freshwater Biology*, *48*, 1311–1328.

Potapova, M. G., & Charles, D. F. (2002). Benthic diatoms in USA rivers: Distributions along spatial and environmental gradients. *Journal of Biogeography*, *29*, 167–187.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Rouse, W. R., Douglas, M. S. V., Hecky, R. E., Hersley, A. E., Kling, G. W., Lesack, L., … Smol, J. P. (1997). Effects of climate change on the freshwaters on arctic and subarctic North America. *Hydrological Processes*, *11*, 873–902.

Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, *7*, 206–218.

Shipley, B. (2009). Confirmatory path analysis in a generalized multilevel context. *Ecology*, *90*, 363–368.

Smol, J. P., & Stoermer, E. F. (2010). *The diatoms: Applications for the environmental and earth sciences*. New York: Cambridge University Press.

Soininen, J., Jamoneau, A., Rosebery, J., & Passy, S. I. (2016). Global patterns and drivers of species and trait composition of diatoms. *Global Ecology and Biogeography*, *25*, 940–950.

Soininen, J., Korhonen, J. J., & Luoto, M. (2013). Stochastic species distributions are driven by organism size. *Ecology*, *94*, 660–670.

Soininen, J., & Luoto, M. (2012). Is catchment productivity a useful predictor of taxa richness in lake plankton communities? *Ecological Applications*, *22*, 624–633.

Soininen, J., Paavola, R., & Muotka, T. (2004). Benthic diatom communities in boreal streams: Community structure in relation to environmental and spatial gradients. *Ecography*, *27*, 330–342.

Sponseller, R. A., Benfield, E. F., & Valett, H. M. (2001). Relationships between land use, spatial scale and stream macroinvertebrate communities. *Freshwater Biology*, *46*, 1409–1424.

Stevenson, R. J. (1997). Scale-dependent determinants and consequences of benthic algal heterogeneity. *Journal of North American Benthological Society*, *16*, 248–262.

Stevenson, R. J., Bothwell, M. L., & Lowe, R. L. (1996). *Algal ecology: Freshwater benthic ecosystems*. San Diego: Elsevier.

Tedersoo, L., Bahram, M., Põlme, S. Kõljalg, U., Yorou, N. S., Wijesundera, R., … Abarenkov, K. (2014) Global diversity and geography of soil fungi. *Science*, *346*, 1078 and 1256688/1–1256688/10.

Tonkin, J. D., Heino, J., Sundermann, A., Haase, P., & Jähnig, S. C. (2016). Context dependency in biodiversity patterns of central German stream metacommunities. *Freshwater Biology*, *61*, 607–620.

Venäläinen, A., & Heikinheimo, M. (2002). Meteorological data for agricultural applications. *Physics and Chemistry of the Earth*, *27*, 1045–1050.

Verleyen, E., Vyverman, W., Sterken, M., Hodgson, D. A., De Wever, A., Juggins, S., … Sabbe, K. (2009). The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos*, *118*, 1239–1249.

Vyverman, W., Verleyen, E., Sabbe, K., Vanhoutte, K., Sterken, M., Hodgson, D. A., … De Wever, A. (2007). Historical processes constrain patterns in global diatom diversitymen. *Ecology*, *88*, 1924–1931.

Wang, J., Price, K. P., & Rich, P. M. (2001). Spatial patterns of NDVI in response to precipitation and temperature in the central Great Plains. *International Journal of Remote Sensing*, *22*, 3827–3844.

Wang, J., Soininen, J., He, J., & Shen, J. (2012). Phylogenetic clustering increases with elevation for microbes. *Environmental Microbiology Reports*, *4*, 217–226.

Weckström, J., Korhola, A., & Blom, T. (1997). The relationship between diatoms and water temperature in thirty subarctic Fennoscandian lakes. *Arctic and Alpine Research*, *29*, 75–92.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, *21*, 213–251.

Winterbourn, M. J., & Collier, K. J. (1987). Distribution of benthic invertebrates in acid, brown water streams in the South Island of New Zealand. *Hydrobiologia*, *153*, 277–286.

Woodcock, S., Van Der Gast, C. J., Bell, T., Lunn, M., Curtis, T. P., Head, I. M., & Sloan, W. T. (2007). Neutral assembly of bacterial communities. *FEMS Microbial Ecology*, *62*, 171–180.

BIOSKETCHES

Virpi Pajunen is a PhD student in physical geography at the University of Helsinki. The focus of her thesis is in species distributions of benthic diatoms in streams. **Janne Soininen** is an associate professor in spatial environmental research at the University of Helsinki. He is interested in large-scale community ecology and especially in the distribution of small aquatic organisms. **Miska Luoto** is a professor at the Department of Geosciences and Geography, University of Helsinki. His study interests are related to the integration of remote sensing and geographical information data in global change modelling.

DATA ACCESSIBILITY

The data will be archived in Dryad.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Pajunen V, Luoto M, Soininen J. Unravelling direct and indirect effects of hierarchical factors driving microbial stream communities. *J Biogeogr*. 2017;44:2376–2385. https://doi.org/10.1111/jbi.13046

Paper IV

Pajunen, V., Luoto, M., Soininen, J. 2016. Stream diatom assemblages as predictors of climate. *Freshwater Biology* 61, 876-886.

Freshwater Biology

Freshwater Biology (2016) **61,** 876–886 doi:10.1111/fwb.12750

Stream diatom assemblages as predictors of climate

VIRPI PAJUNEN, MISKA LUOTO AND JANNE SOININEN
Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland

SUMMARY

1. Benthic diatoms have been widely used as indicators of water quality in streams. New insights that diatoms may also respond to large-scale drivers, such as climate or historical factors, highlight the need to reassess the usefulness and the reliability of diatoms as bioindicators.
2. Using a suite of modelling techniques, weighted averaging (WA), weighted averaging partial least squares, modern-analogue technique (MAT) and two machine learning techniques, boosted regression trees (BRT) and random forests (RF), we calibrated models to infer water quality and climatic variables using diatom abundance data collected from 227 stream sites in Finland.
3. Predictive ability was generally better for climatic variables [growing degree days (GDD) defined as temperature >5 °C, summer precipitation and water balance] than for local environmental variables (conductivity, water colour and total phosphorus). The strongest relationships were found for GDD (r^2 = 0.86, MAT) and conductivity (r^2 = 0.82, RF). Using BRT, we also identified potential indicator species for local environmental and climatic variables, based on relative importance species in the models.
4. Our results show that diatoms could serve as efficient proxies for climatic variables and local environmental conditions. Furthermore, new modelling techniques such as modern regression trees can provide new insights into relationships between diatom assemblages and local water quality and climate, and thus help to construct more reliable indices. These methods could also serve as important tools to infer environmental variables in changing ecosystems.

Keywords: bioindicators, biomonitoring tools, climate, regression tree methods, stream diatoms

Introduction

Diatoms are considered to be ubiquitous, like many other microbes, with their distributions largely governed by local environmental variables, mainly water chemistry (Finlay, 2002; Soininen, Paavola & Muotka, 2004). An increasing number of studies have shown that, not only local variables, but also large-scale factors related to climate (Vyverman & Sabbe, 1995; Weckström, Korhola & Blom, 1997; Leira & Sabater, 2005; Berthon *et al.*, 2014), history (Vyverman *et al.*, 2007) and dispersal (Verleyen *et al.*, 2009) influence diatom species distribution. The effects of these factors are related to spatial scale, with stronger effects at broad geographical scales, while local environmental variables become more predominant at smaller, regional scales (Martiny *et al.*, 2006; Bennett *et al.*, 2010; Astorga *et al.*, 2012).

Regional factors affect diatom assemblages through indirect pathways that are often related to resources and environmental stress (Stevenson, 1997; Rühland, Paterson & Smol, 2015). For example, hydrological and thermal regimes affect seasonality, flow conditions, runoff and the rate of primary production and metabolism (Allan & Castillo, 2007). Accordingly, benthic diatoms adapt to prevailing conditions, by finding suitable habitat, through efficient nutrient uptake and by their ability to tolerate drought, low light and chemical conditions. Hence, the effects of large-scale drivers on diatom assemblages is complex and deserves further study, not only because the current use of diatoms in biomonitoring is based on their responses to physicochemical properties that are typically are highly variable (Bottin *et al.*, 2014) but also because of changing climate. The projected climate change can alter the thermal regimes in streams by increasing in-stream temperatures directly

Correspondence: Virpi Pajunen, Department of Geosciences and Geography, University of Helsinki, P.O. Box 64, Helsinki FI-00014, Finland. E-mail: virpi.pajunen@helsinki.fi.

(Isaak *et al.*, 2010) and indirectly via impacts on flow regimes through increasing precipitation or drought and decreasing snow cover (IPCC, 2013). As climatic factors can affect diatom assemblages indirectly via several local environmental factors (Stevenson, 1997), these factors could strengthen or alter the effect of local environmental variables, thus affecting the responses of indicator species (Rühland *et al.*, 2015). Here, we suggest that new methodological approaches based on the distribution of individual species may provide novel insights into diatom distribution along environmental and climatic gradients.

There has been a long tradition of using diatoms as indicators of past climatic conditions in palaeolimnological studies, and more recently as indicators of changes in mean annual temperature (reviewed in Smol, 2010). Species-specific and community models have been used to infer past environmental conditions by constructing either quantitative statistical models, such as weighted averaging (WA) and weighted averaging partial least squares (WA-PLS), or models based on modern-analogue matching (i.e. modern-analogue technique, MAT) (Birks *et al.*, 2010). Overall, these methods are relatively good at modelling biological response variables (explanatory and predictive power), although WA cannot identify more complex response shapes (such as thresholds) (De'ath, 2007).

Recently, machine learning techniques, such as boosted regression trees (BRT) and random forests (RF), have become more popular in ecological studies (e.g. Leathwick *et al.*, 2006; Cutler *et al.*, 2007; Lewin *et al.*, 2014) and BRTs, in particular, have been used successfully in reconstructing palaeoclimate using pollen assemblages (Salonen *et al.*, 2014). BRTs and RFs have been shown to have many advantages, such as lower prediction error, when compared to other methods such as GLM and GAM (Elith *et al.*, 2006; Leathwick *et al.*, 2006; Cutler *et al.*, 2007; De'ath, 2007). Moreover, these regression tree methods: (i) can manage various types of predictor variables, (ii) do not require prior data transformation, (iii) are able to fit complex nonlinear relationships, (iv) automatically take into account interaction effects between predictors (Cutler *et al.*, 2007; Elith, Leathwick & Hastie, 2008) and (v) show the relative importance of each predictive variable (Cutler *et al.*, 2007; De'ath, 2007).

In this study, our aim was to determine the efficacy of diatom assemblages for predicting climatic [growing degree days (GDD), precipitation and water balance] and local environmental [total phosphorus (TP), conductivity and water colour] variables using diatom assemblages from 227 streams and five calibration models: WA, WA-PLS, MAT, BRT and RF. We were particularly interested to determine if stream diatoms could be utilised as efficient proxies for local environmental conditions as well as for large-scale environmental factors related to climate. Finally, using BRT, we identified indicator species for selected local stream characteristics and climatic variables.

Methods

Data collection

An extensive data set of stream diatom assemblages from 227 sites was obtained by combining data from three different studies done in Finland (60°–70° N, 20°–32° E) (Fig. 1). Between 1996 and 2001, diatoms were sampled from 141 sites, comprising the five ecoregions in Finland and covering broad gradients in conductivity, pH, humic content and nutrient concentrations (Soininen *et al.*, 2004) (see Appendix Table S1 in Supporting Information). Sites are described in detail by Soininen (2002) and Soininen *et al.* (2004). The second data set comprised 56 fast-flowing rivers in central Finland sampled in 1986 by Eloranta (1995). These sites were included because most of them represent near-pristine conditions, that is, only slightly affected by agriculture and fish farming. The third data set comprised 30 sites sampled in July 2004 in northern Lapland. We consider all samples to be comparable as they were sampled using the same methods. In addition, Korhonen, Köngäs & Soininen (2013) demonstrated that diatom assemblages are robust indicators of water chemistry variables even though assemblages vary in time.

Diatom sampling

Diatoms were sampled by brushing stones with a toothbrush, according to the recommendations of Kelly *et al.* (1998). At least five replicate, pebble-to-cobble sized stones (5–15 cm) were selected randomly from five to 10 cross-stream transects, and brushed for diatoms. The diatom suspension was placed in a small plastic bottle and preserved in ethanol (70%). All sampling was conducted during low flow conditions in July and August. In the laboratory, diatom samples were cleaned from organic material by wet combustion with acid (HNO3 : H2SO4; 2 : 1) and mounted in Naphrax or Dirax. A total of 250–500 valves per sample were identified and counted using phase contrast light microscopy (magnification 1000×). Species were identified according

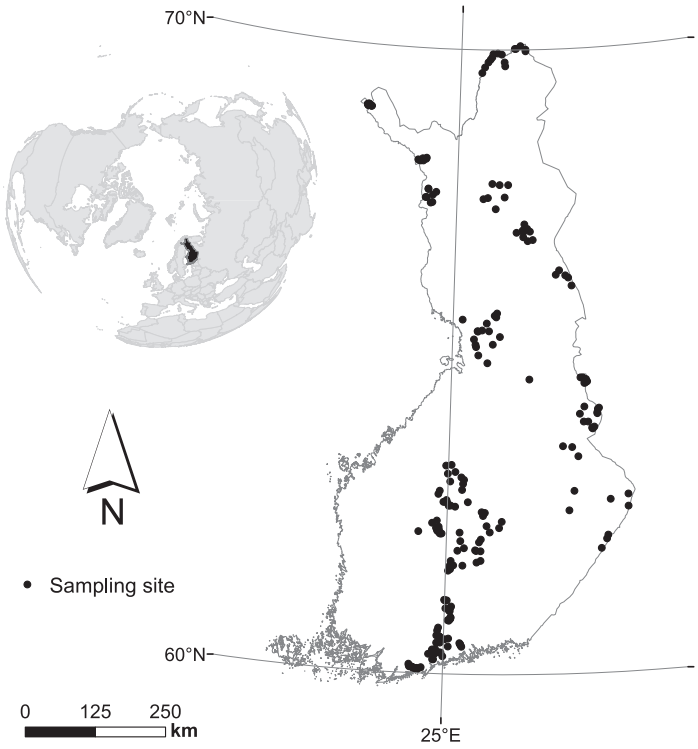


Fig. 1 Location of the sampling sites (n = 227) in Finland, northern Europe.

to Krammer & Lange-Bertalot (1986–1991) and Lange-Bertalot & Metzeltin (1996) by two analysts who harmonised species identification.

Environmental variables

At most of the sites, water samples were taken simultaneously with the diatom samples and analysed for TP, pH, conductivity and water colour. For less than 20% of the sites, water chemistry data were taken from the national water quality database, using results from the nearest sampling occasion in space and time. Current velocity, shading by the riparian vegetation and stream width were measured at each site along 10 transects perpendicular to the flow covering the whole sampling area.

Climate data (averages for the years 1981–2010) were obtained from Finnish Meteorological Institute. As water temperature can fluctuate strongly even at relatively small temporal scales (Allan & Castillo, 2007), especially in small streams, we used GDD (defined as temperature >5 °C) as a proxy for overall productivity of the stream and its catchment. Summer precipitation (sum from May to September; PRECS) and water balance (WAB) were used as measures of atmospheric water supply to the whole catchment area and potential runoff, respectively. WAB was calculated according to Skov & Svenning (2004) by summing the monthly differences between precipitation and potential evapotranspiration (PET). Monthly PET was calculated as

$$\mathrm{PET} = 58.93 \times T_{\mathrm{bio}}/12 \quad (1)$$

where T_{bio} is the Holdridge biotemperature, defined as the annual mean of monthly temperatures with negative monthly values adjusted to zero (Holdridge, 1967; Lugo *et al.*, 1999). Multiple linear regression was used to relate climate data to latitude, longitude and altitude of each study site, downscaling climate data from 10 × 10-km resolution grid to the study site (Finnish Meteorological Institute; Venäläinen & Heikinheimo, 2002).

Principal component analysis (PCA), redundancy analysis (RDA) and partial redundancy analysis (pRDA) were used to correlatively explore interrelations between

climatic and local environmental variables and between environmental variables and diatom abundance. In PCA, the first two components collectively explained 57.6% of variation in the original variables. GDD, mean January temperature (T_{Jan}), TP and conductivity had high positive loadings on the first principal component, while pH and conductivity had high positive loadings and precipitation and water colour had high negative loadings on the second component (see Fig. S1 and Table S2). Covariance between the local and climatic variables was assessed using Spearman's rank correlation. The predictor variables only exhibited moderate collinearity (maximum pairwise $r_s = 0.65$; all other correlations $\leq|0.63|$), and therefore all of the variables were retained for subsequent modelling (see Fig. S2).

Based on results from PCA and RDA, we selected TP, conductivity and water colour to represent local environmental variables, and GDD, PRECS and WAB to represent climatic variables (see Table S3 and Fig. S3). Variation partitioning confirmed the importance of these variables as strong predictors of diatom abundance; joint effects explained most of the explained total variation in both partitions (10% with full set of variables and 8% with six variables) (see Fig. S4). The explained variation in local environment (8 and 7%) was higher than for climatic variables (5%). PCA, RDA and pRDA were performed in R (version 3.1.1; R Development Core Team, 2014) applying VEGAN package (Oksanen *et al*., 2015).

Inference models

Five different modelling approaches were used in calibration: WA, WA-PLS, MAT, BRT and RF. In model calibration, conductivity, TP, water colour, GDD, PRECS and WAB were set as response variables and the diatom abundance of 214 taxa from 227 sites as predictors. All models were fitted using R software (version 3.1.1; R Development Core Team, 2014).

The WA, WA-PLS and MAT models were run using functions from the RIOJA package version 0.8–5 (Juggins, 2013b) on square-root transformed species data (Prentice, 1980). The WA technique is based on the assumption of unimodal relationships between species and environmental variables (ter Braak & van Dam, 1989). WA-PLS involves a weighted inverse deshrinking regression (ter Braak & Juggins, 1993) and generally shows improved performance compared with WA (Birks *et al*., 2010). WA was run using monotonic deshrinking, while WA-PLS was run using three-component models. In MAT, an analogue is compared numerically to species abundance data using a measure of dissimilarity. MAT was based on the weighted mean of the k closest analogues and squared chord distance (Overpeck, Webb & Prentice, 1985). The screenplot function (R package ANALOGUE version 0.10–0; Simpson & Oksanen, 2013) was used to determine the optimal values for k. The k values used for the response variables were 5.

The machine learning techniques BRT and RF were also used to model diatom abundance. In BRTs, predictions with minimised loss function (such as deviance) are composed by a boosting method in which a sequence of simple regression trees is gradually grown fitting one tree at the time to the sequence (Friedman, Hastie & Tibshirani, 2000; De'ath, 2007; Elith *et al*., 2008). The BRT model was fitted using functions from the GBM package version 1.6–3.1 (Ridgeway, 2014), which is based on the Gradient Boosting Machine developed by Friedman (2001). The interaction depth in the model was set to 6. The learning rate, that determines the contribution of each tree to the growing model, was set to 0.005. The maximum number of trees was set to 1000. A Gaussian distribution of errors was used to model the six variables. In RF, designed to produce accurate predictions that do not overfit the data, a large number of trees (i.e. a forest) are grown with a randomised subset of predictors (Breiman, 2001). The RF model was run using the functions from the R package randomForest. The number of trees (k) was set to 500 and the minimum size of terminal nodes was set to 5.

Leave-one-out cross-validation was used to assess the performance of all five models. Model performance was estimated by the coefficient of determination (r^2) and the root-mean-square error of prediction (RMSEP). In the BRT model, the relative influence of the predictor variables was first estimated according to Friedman (2001) and then scaled to sum up to 100. The higher value a predictor variable gets, the stronger its influence on the response variable.

Spatial autocorrelation, which may affect significance and confidence levels of model estimates (Legendre *et al*., 2002), was evaluated for each environmental variable by generating spatial correlograms (R package pgirness) using raw data and model residuals. Moran's I coefficients were calculated for 10 distance class intervals. Significant values of Moran's I indicate that pairs of localities within a given distance are either similar (positive values) or dissimilar (negative values) (Legendre & Legendre, 1998). As a correlogram visualises the level of autocorrelation as a function of spatial distance, it also describes the level of spatial dependence and the shape of spatial structure of each environmental variable in the data. The significance level of each Mor-

an's I coefficient was evaluated with 999 permutations (Sokal & Oden, 1978a,b). A correlogram was considered to be significant ($P \leq 0.05$) if at least one of its coefficients was significant at P/k, where k is the number of distance classes used (following the Bonferroni criterion). Spatial autocorrelation was substantially smaller in model residuals compared to raw data (see Fig. S5).

Results

Of the local environmental variables, predictive performance was the highest for conductivity; coefficients of determination between observed and diatom inferred were >0.75 in all predictive models (RF, $r^2 = 0.82$; BRT, $r^2 = 0.80$; WA-PLS, $r^2 = 0.80$; MAT, $r^2 = 0.79$; WA, $r^2 = 0.75$) (Fig. 2). Predictive performance was lower for water colour [r^2 values ranged from 0.34 (WA) to 0.52 (RF)] and for TP [r^2 values from 0.54 (MAT) to 0.60 (WA and WA-PLS)] in all models (Table 1, see Fig. S6).

Regression tree methods, BRT and RF, had the best overall predictive performance. Both regression tree methods and MAT models were, on average, better in predicting climatic variables than the local environmental variables (Fig. 3). However, in WA and WA-PLS models the average predictive performance was similar for both local environmental and climatic variables. For climatic variables, the predictive performance was best for GDD (r^2 values ranged from 0.67 for WA to 0.86 for MAT) (Table 1). For WAB, r^2 values ranged from 0.34 (WA) to 0.73 (RF) and for PRECS, r^2 values ranged from 0.56 (WA) to 0.75 (BRT). There were differences in the predictive performances (r^2) of the models between the variable groups; WA and WA-PLS were best at predicting TP, while MAT and RF were best at predicting the climatic variables. Prediction errors (RMSEP) were typically highest in WA and lowest in BRT and RF.

For all six predicted response variables, three to five species were consistently found with a relative importance >5% based on BRT (Table 2). For TP, *Nitzschia palea* was found to have the highest relative importance (27%). Species that had high relative importance for conductivity predictions were *Surirella brebissonii* (40.2%) and *Nitzschia levidensis* (10.6%). The abundances of *N. levidensis* and *S. brebissonii* increased with conductivity (see Fig. S7). For water colour, the species with highest relative importance was *Eunotia meisteri* (14.4%). For GDD, the species with the highest relative importance were *Achnanthes pusilla* (31.8%) and *Caloneis tenuis* (10.6%), which showed high sensitivity to the length of the growing season (see Fig. S7). For PRECS, the most important species were *A. pusilla* (18.6%) and *Fragilaria*

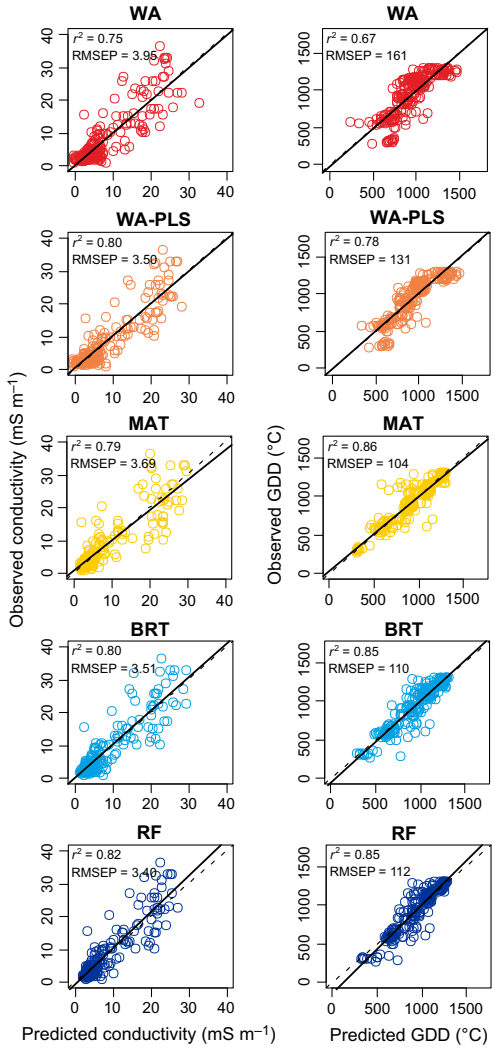


Fig. 2 Relationships between observed and diatom inferred values for conductivity and growing degree days (GDD) using five different calibration models: weighted averaging (WA), weighted averaging partial least squares (WA-PLS), modern-analogue technique (MAT), boosted regression trees (BRT) and random forests (RF). Each plot shows the coefficient of determination (r^2) and root-mean-square error of prediction (RMSEP) and fitted linear models presented as dashed lines. In WA and WA-PLS plots, linear models have a slope very close to 1.

capucina var. *gracilis* (14.2%). Finally, the species with highest relative importance for WAB predictions was *Fragilaria exigua* (9.6%).

Table 1 The coefficient of determination (r^2) and root-mean-square error of prediction (RMSEP) for five calibration models [weighted averaging (WA), weighted averaging partial least squares (WA-PLS), modern-analogue technique (MAT) boosted regression trees (BRT) and random forests (RF)] used to infer local environmental [total phosphorus (TP), conductivity and water colour] and climatic [growing degree days (GDD), summer precipitation sum from May to September (PRECS) and water balance (WAB)] variables with diatom abundance data.

	WA		WA-PLS		MAT		BRT		RF	
Response variable	r^2	RMSEP	r^2	RMSEP	r^2	RMSEP	r^2	RMSEP	r^2	RMSEP
TP (μg L^{-1})	0.60	20.2	0.60	20.2	0.54	22.0	0.54	21.6	0.58	20.9
Cond. (mS/m)	0.75	3.95	0.80	3.50	0.79	3.69	0.80	3.51	0.82	3.40
Colour (mg Pt L^{-1})	0.34	58	0.48	51.1	0.44	54	0.47	52	0.52	50
GDD (°C)	0.67	161	0.78	131	0.86	104	0.85	110	0.85	112
PRECS (mm)	0.56	16.7	0.66	14.7	0.74	12.8	0.75	12.6	0.74	13.4
WAB (mm)	0.34	31.4	0.45	28.5	0.72	20.5	0.68	21.9	0.73	21.3

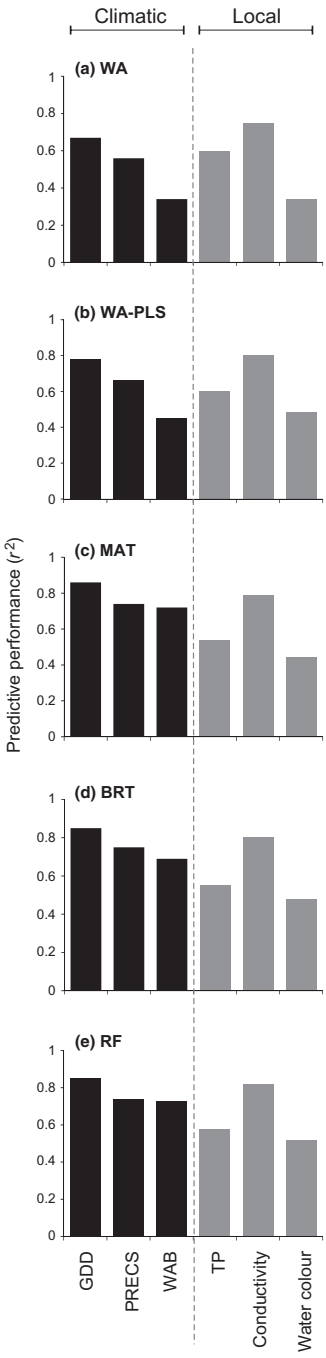
Discussion

In bioassessment it is often desirable to use methods that reflect abiotic conditions over extended periods of time, that is, longer than the snapshot measure of a water chemistry sample. Our results show that relative abundances of diatoms can be reliably used to predict not only local environmental conditions in streams but also climatic variables. Indeed, one of the most important findings of our study was that we could predict climatic variables using diatoms more accurately than local environmental variables. This finding shows that diatoms respond strongly to large-scale climatic variables (Vyverman & Sabbe, 1995; Weckström *et al*., 1997; Leira & Sabater, 2005) and not only local factors (Vyverman *et al*., 2007). In other words, although local environmental conditions (i.e. water chemistry and physical factors) may govern diatom species distribution locally (Verleyen *et al*., 2003; Soininen, 2007), distributions are also affected by climate-driven processes via multiple pathways (Stevenson, 1997). This finding implies that climatic variables reflect the influences of latent local environmental variables that covary.

The observed high predictive performance for climatic variables, compared to measures of stream physiochemistry, shows that long-term climatic data are robust in predicting diatom abundance. As water chemistry and diatoms were collected at the same time, this finding suggests that diatoms are responding more to mean water chemistry than in-stream conditions at the time of sampling. Using snapshot measures of water chemistry might therefore increase model uncertainty. Our study suggests that climatic variables that can be easily extracted from extensive data bases were found to be robust and convenient proxy variables for more complex models. However, a mechanistic understanding of the species–environment relationship is needed to avoid unreliable and misleading models (Juggins, 2013a).

Models for GDD had the highest predictive performance, indicating the importance of temperature-related processes (e.g. productivity) on diatom abundances. *Achnanthes pusilla* and *C. tenuis* were strongly (inversely) associated with GDD (see Fig. S7). Classified as nitrogen-autotrophs, these two species prefer habitats with low levels of organic nitrogen, good water quality and high oxygen concentrations (van Dam, Mertens & Sinkeldam, 1994). Studies on lakes in subarctic regions, where fresh water is typically cold and oligotrophic, have also indicated that *A. pusilla* has a low temperature optimum (Weckström *et al*., 1997; Bigler & Hall, 2002). Furthermore, Vyverman & Sabbe (1995) observed that the occurrence of *C. tenuis* in the tropics was confined to high altitude lakes, suggesting a preference for relatively cold, harsh environmental conditions. The relatively strong relationships between *A. pusilla* and *C. tenuis* and GDD could be best explained by their responses to organic material; GDD affects primary production which increases organically bound nutrients. Enhanced primary production also increases the amount of organic material and its decomposition consumes oxygen, with decomposition accelerated at increased temperatures (Allan & Castillo, 2007). Thus, GDD can affect benthic diatoms via multiple pathways (e.g. overall in-stream or catchment productivity, nutrient concentrations, amount of organic material and dissolved oxygen and water temperature).

The effect of climate on stream diatoms was further emphasised by our finding that diatom- based models for summer precipitation and WAB were robust, especially when using RF, BRT or MAT. Precipitation may affect diatom abundances through hydrology and associated changes in water chemistry. The species that had



the highest relative importance for PRECS (*A. pusilla* and *F. capucina* var. *gracilis*) and WAB (*F. exigua*) are ecologically similar, with a preference for harsh, oxygen-rich, low-nutrient conditions (van Dam *et al.*, 1994). *Achnanthes pusilla* and *F. capucina* var. *gracilis* were most abundant at sites with lowest sum of summer precipitation, whereas *F. exigua* exhibited a more complex response to WAB. Hydrological factors, such as current velocity and flow regime, are known to influence stream diatom assemblages via disturbance (Rott *et al.*, 2006; Passy, 2007). *F. capucina* var. *gracilis* has been classified as a high profile species, sensitive to high current velocities and thus abundant in low current velocities (Passy, 2007). In our data, both *A. pusilla* and *F. capucina* var. *gracilis* were most abundant at sites with moderate current velocity (20–70 cm s^{-1}). Increased precipitation may also impair water quality by increasing nutrient loads and fine sediments from the catchment (Allan & Castillo, 2007). However, precipitation can also have a dilutive effect, depending on the characteristics of the catchment.

In agreement with previous work, our study identified a number of species responding to local environmental conditions. For example, from the literature a number of species have been identified as indicators of eutrophic (*N. palea*, *Melosira varians*, *S. brebissonii*) and mesotrophic (*Fragilaria capucina*, *Gomphonema parvulum*) conditions (van Dam *et al.*, 1994; Fore & Grafe, 2002; Rimet *et al.*, 2005). In our study, *Surirella brebissonii* and *M. varians* showed a preference for both high conductivity and phosphorus concentrations, confirming their use as indicators of pollution. Together with *Navicula gregaria*, *S. brebissonii* and *M. varians* have also been recognised as indicators for high conductivity (Potapova & Charles, 2003; Urrea & Sabater, 2009). *E. meisteri* showed a preference for humic conditions; abundant at sites with water colour >80 mg Pt L^{-1}.

Stream diatoms have previously been used to infer nutrient concentrations using WA (e.g. Winter & Duthie, 2000; Soininen & Niemelä, 2002). Our findings confirm that WA is a superior technique when modelling TP concentrations. However, WA assumes a unimodal relationship between species' abundances and environmental, which is often too simplistic of an assumption as

Fig. 3 Predictive performance (r^2) for diatom-based weighted averaging (WA), weighted averaging partial least squares (WA-PLS), modern-analogue technique (MAT), boosted regression trees (BRT) and random forests (RF) models for climatic and local environmental variables.

Table 2 The five most important predictors (diatom taxa) and their relative importance in BRT for the local environmental [total phosphorus (TP), conductivity and water colour] and climatic [growing degree days (GDD), summer precipitation sum from May to September (PRECS) and water balance (WAB)] variables in streams.

Response variable	Taxon name	Relative importance	Maximum %	Mean %	Prevalence
TP	*Nitzschia palea*	27.0	21.0	1.4	0.47
TP	*Melosira varians*	9.2	71.5	1.5	0.21
TP	*Surirella brebissonii*	6.3	44.7	1.0	0.23
TP	*Gomphonema parvulum*	3.6	31.2	2.4	0.64
TP	*Fragilarica capucina*	3.5	43.6	4.4	0.62
Cond.	*Surirella brebissonii*	40.2	44.7	1.0	0.23
Cond.	*Nitzschia levidensis*	10.6	11.4	0.2	0.13
Cond.	*Navicula gregaria*	9.0	12.9	0.4	0.15
Cond.	*Tabellaria flocculosa*	7.7	69.4	7.4	0.88
Cond.	*Melosira varians*	4.8	71.5	1.5	0.21
Colour	*Eunotia meisteri*	14.4	11.2	0.7	0.33
Colour	*Fragilaria arcus*	8.2	32.1	1.0	0.28
Colour	*Eunotia implicata*	5.9	19.0	0.7	0.30
Colour	*Achnanthes minutissima*	5.6	75.0	17.7	0.96
Colour	*Eunotia bilunaris*	4.8	30.8	1.5	0.70
GDD	*Achnanthes pusilla*	31.8	44.9	1.4	0.30
GDD	*Caloneis tenuis*	10.6	2.6	0.1	0.17
GDD	*Eunotia implicata*	5.9	19.0	0.7	0.30
GDD	*Gomphonema gracile*	4.2	9.1	0.3	0.33
GDD	*Eunotia arcus*	3.7	3.2	0.1	0.13
PRECS	*Achnanthes pusilla*	18.6	44.9	1.4	0.30
PRECS	*Fragilaria capucina* var. *gracilis*	14.2	47.4	0.8	0.11
PRECS	*Caloneis tenuis*	7.1	2.6	0.1	0.17
PRECS	*Achnanthes linearis*	6.3	24.4	1.5	0.56
PRECS	*Navicula heimansioides*	5.9	5.2	0.2	0.11
WAB	*Fragilaria exigua*	9.6	3.2	0.1	0.10
WAB	*Eunotia rhomboides*	7.0	15.3	0.2	0.12
WAB	*Navicula heimansioides*	6.1	5.2	0.2	0.11
WAB	*Achnanthes biasolettiana*	5.1	12.8	0.1	0.13
WAB	*Navicula cocconeiformis*	3.8	2.0	<0.1	0.14

thresholds for species' occurrences are common in fresh water (Potapova *et al.*, 2004; Soininen, Korhonen & Luoto, 2013). As also shown in our study, diatom species have clear thresholds along environmental or climatic gradients. In our study, MAT and the regression tree methods out-performed WA and WA-PLS when modelling climatic variables. This might be due to the ability of these modelling techniques to account for complex nonlinear relationships related to large-scale climatic variables. Compared to MAT and both WA methods, BRT and RF had consistently good predictive performance and lower prediction errors for most of the inferred variables. Therefore, they could be regarded as potentially superior methods compared to WA, WA-PLS and MAT.

From an applied perspective, our results emphasise the need to re-evaluate the accuracy of diatom indices currently used in biomonitoring, as large-scale climatic factors may have a strong influence on stream diatoms via multiple local variables. Moreover, species responses to environmental factors can be nonlinear and, therefore, methods that are able to recognise more complex relationships are needed in the development of new indices. The reliability of water quality is directly related to its ability to reflect the prevailing local environmental conditions. Some researchers have already questioned the use of diatom indices in biomonitoring due to large discrepancies between geographical regions (e.g. in the trophic value scores and resilience of the communities after environmental changes) (Rimet *et al.*, 2005; Pota-

pova & Charles, 2007; Coste *et al.*, 2009; Besse-Lototskaya *et al.*, 2011). To highlight the importance of our results further, the methods used here might bring novel applications for palaeoecology and climate change research, for example, to reconstruct past climatic conditions. Until now, the amount of fluvial diatom species in sediments has been used to hindcast runoff. However, the ability to successfully predict climatic variables from stream diatom data could enable the inference of past climatic conditions from fossil diatom assemblages delivered from, for example, sediments of ancient stream deltas.

Acknowledgments

This project was funded by Maj and Tor Nessling Foundation. We thank two anonymous reviewers for the insightful comments that greatly improved the manuscript.

References

Allan J.D. & Castillo M.M. (2007) *Stream Ecology: Structure and Function on Running Waters*, Second Edition. Springer, Dordrecht.

Astorga A., Oksanen J., Luoto M., Soininen J., Virtanen R. & Muotka T. (2012) Distance decay of similarity in freshwater communities: do macro- and microorganisms follow the same rules? *Global Ecology and Biogeography*, **21**, 365–375.

Bennett J.R., Cumming B.F., Ginn B.K. & Smol J.P. (2010) Broad-scale environmental response and niche conservatism in lacustrine diatom communities. *Global Ecology and Biogeography*, **19**, 724–732.

Berthon V., Alric B., Rimet R. & Perga M.-E. (2014) Sensitivity and responses of diatoms to climate warming in lakes heavily influenced by humans. *Freshwater Biology*, **59**, 1755–1767.

Besse-Lototskaya A., Verdonschot P.F.M., Coste M. & van de Vijver B. (2011) Evaluation of European diatom trophic indices. *Ecological Indicators*, **11**, 456–467.

Bigler C. & Hall R. (2002) Diatoms as indicators of climatic and limnological change in Swedish Lapland: a 100-lake calibration set and its validation for paleoecological reconstructions. *Journal of Paleolimnology*, **27**, 97–115.

Birks H.J.B., Heiri O., Seppä H. & Bjune A.E. (2010) Strengths and weaknesses of quantitative climate reconstructions based on Late-Quaternary biological proxies. *The Open Ecology Journal*, **3**, 68–110.

Bottin M., Soininen J., Ferrol M. & Tison-Rosebery J. (2014) Do spatial patterns of benthic diatom assemblages vary across regions and years? *Freshwater Science*, **33**, 402–416.

Breiman L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Coste M., Boutry S., Tison-Rosebery J. & Delmas F. (2009) Improvements of the Biological Diatom Index (BDI): description and efficiency of the new version (BDI-2006). *Ecological Indicators*, **9**, 621–650.

Cutler D.R., Edwards T.C., Beard K.H., Cutler A. & Hess K.T. (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

De'ath G. (2007) Boosted trees for ecological modeling and prediction. *Ecology*, **88**, 243–251.

Elith J., Graham C., Anderson R., Dudik M., Ferrier S., Guisan A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith J., Leathwick J.R. & Hastie T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Eloranta P. (1995) Type and quality of river waters in central Finland described using diatom indices. In: *Proceedings of the 13th International Diatom Symposium* (Eds D. Marino & M. Montresor), pp. 271–280. Biopress, Bristol.

Finlay B. (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Fore L. & Grafe C. (2002) Using diatoms to assess the biological condition of large rivers in Idaho (USA). *Freshwater Biology*, **47**, 2015–2037.

Friedman J., Hastie T. & Tibshirani R. (2000) Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, **28**, 337–407.

Friedman J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

Holdridge L.R. (1967) *Life Zone Ecology*. Tropical Science Center, Santa Jose, CA.

IPCC (2013) Summary for policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Eds T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P.M. Midgley), pp 1–29, Cambridge University Press, Cambridge.

Isaak D.J., Luce C.H., Rieman B.E., Nagel D.E., Peterson E.E., Horan D.L. *et al.* (2010) Effects of climate change and wildfire on stream temperatures and salmonid thermal habitat in a mountain river network. *Ecological Applications*, **20**, 1350–1371.

Juggins S. (2013a) Quantitative reconstructions in palaeolimnology: new paradigm or sick science? *Quaternary Science Reviews*, **64**, 20–32.

Juggins S. (2013b) *Rioja: Analysis of Quaternary Science Data*. http://cran.r-project.org/web/packages/rioja/index.html Accessed online 24 November 2014.

Kelly M., Cazaubon A., Coring E., Dell'Uomo A., Ector L., Goldsmith B. *et al.* (1998) Recommendations for the routine sampling of diatoms for water quality assessments in Europe. *Journal of Applied Phycology*, **10**, 215–224.

Korhonen J.J., Köngäs P. & Soininen J. (2013) Temporal variation of diatom assemblages in oligotrophic and eutrophic streams. *European Journal of Phycology*, **48**, 141–151.

Krammer K. & Lange-Bertalot H. (1986–1991) *Bacillariophyceae. Süßwasserflora von Mitteleuropa*. Fischer, Stuttgart.

Lange-Bertalot H. & Metzeltin D. (1996) Indicators of oligotrophy, 800 taxa representative of three ecologically distinct lake types: carbonate buffered, oligodystrophic, weakly buffered soft water. In: *Iconographia Diatomologica, Vol. 2*, (Ed H. Lange-Bertalot), 390 pp. Koeltz Scientific Books, Königstein.

Leathwick J.R., Elith J., Francis M.P., Hastie T. & Taylor P. (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Marine Ecology Progress Series*, **321**, 267–281.

Legendre P., Dale M., Fortin M.-J., Gurevitch J., Hohn M. & Myers D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. *Ecography*, **25**, 601–615.

Legendre P. & Legendre L. (1998) *Numerical Ecology*, 2nd edn. Elsevier Science, Amsterdam.

Leira M. & Sabater S. (2005) Diatom assemblages distribution in catalan rivers, NE Spain, in relation to chemical and physiographical factors. *Water Research*, **39**, 73–82.

Lewin W.-C., Mehner T., Ritterbusch D. & Braemick U. (2014) The influence of anthropogenic shoreline changes on the littoral abundance of fish species in German lowland lakes varying in depth as determined by boosted regression trees. *Hydrobiologia*, **724**, 293–306.

Lugo A., Brown S., Dodson R., Smith T. & Shugart H. (1999) The Holdridge life zones of the conterminous United States in relation to ecosystem mapping. *Journal of Biogeography*, **26**, 1025–1038.

Martiny J.B.H., Bohannan J.M., Brown J.H., Colwell R.K., Fuhrman J.A., Green J.L. *et al*. (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102–112.

Oksanen J., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B. *et al*. (2015). *Vegan: Community Ecology Package*. Available at: http://cran.r-project.org/web/packages/vegan/index.html (accessed June 2015).

Overpeck J.T., Webb T. III & Prentice I.C. (1985) Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quaternary Research*, **23**, 87–108.

Passy S.I. (2007) Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters. *Aquatic Botany*, **86**, 171–178.

Potapova M., Charles D., Ponader K. & Winter D. (2004) Quantifying species indicator values for trophic diatom indices: a comparison of approaches. *Hydrobiologia*, **517**, 25–41.

Potapova M. & Charles D.F. (2003) Distribution of benthic diatoms in US rivers in relation to conductivity and ionic composition. *Freshwater Biology*, **48**, 1311–1328.

Potapova M. & Charles D.F. (2007) Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecological Indicators*, **7**, 48–70.

Prentice I.C. (1980) Multidimensional scaling as a research tool in Quaternary palynology – a review of theory and methods. *Review of Palaeobotany and Palynology*, **31**, 71–104.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org.

Ridgeway G. (2014) *gbm: Generalized Boosted Regression Models*. 2010. Accessed online 24 November 2014 at: http://cran.r-project.org/web/packages/gbm/index.html

Rimet F., Cauchie H., Hoffmann L. & Ector L. (2005) Response of diatom indices to simulated water quality improvements in a river. *Journal of Applied Phycology*, **17**, 119–128.

Rott E., Cantonati M., Füreder L. & Pfister P. (2006) Benthic algae in high altitude streams of the Alps – a neglected component of the aquatic biota. *Hydrobiologia*, **562**, 195–216.

Rühland K.M., Paterson A.M. & Smol J.P. (2015) Lake diatom responses to warming: reviewing the evidence. *Journal of Paleolimnology*, **54**, 1–35.

Salonen J.S., Luoto M., Alenius T., Heikkila M., Seppä H., Telford R.J. *et al*. (2014) Reconstructing palaeoclimatic variables from fossil pollen using boosted regression trees: comparison and synthesis with other quantitative reconstruction methods. *Quaternary Science Reviews*, **88**, 69–81.

Simpson G.L. & Oksanen J. (2013) *Analogue: Analogue and Weighted Averaging Methods for Palaeoecology*. http://cran-r-project.org/web/packages/analogue/index.html Accessed online 24 November 2014.

Skov F. & Svenning J. (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. *Ecography*, **27**, 366–380.

Smol J.P. (2010) The power of the past: using sediments to track the effects of multiple stressors on lake ecosystems. *Freshwater Biology*, **55**(Suppl. 1), 43–59.

Soininen J. (2002) Responses of epilithic diatom communities to environmental gradients in some Finnish rivers. *International Review of Hydrobiology*, **87**, 11–24.

Soininen J. (2007) Environmental and spatial control of freshwater diatoms – a review. *Diatom Research*, **22**, 473–490.

Soininen J., Korhonen J.J. & Luoto M. (2013) Stochastic species distributions are driven by organism size. *Ecology*, **94**, 660–670.

Soininen J. & Niemelä P. (2002) Inferring the phosphorus levels of rivers from benthic diatoms using weighted averaging. *Archiv Fur Hydrobiologie*, **154**, 1–18.

Soininen J., Paavola R. & Muotka T. (2004) Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. *Ecography*, **27**, 330–342.

Sokal R.R. & Oden N.L. (1978a) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10**, 199–228.

Sokal R.R. & Oden N.L. (1978b) Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society*, **10**, 229–249.

Stevenson R.J. (1997) Scale-dependent determinants and consequences of benthic algal heterogeneity. *Journal of the North American Benthological Society*, **16**, 248–262.

ter Braak J.F. & Juggins S. (1993) Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia*, **269/270**, 485–502.

ter Braak J.F. & van Dam H. (1989) Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia*, **178**, 209–223.

Urrea G. & Sabater S. (2009) Epilithic diatom assemblages and their relationship to environmental characteristics in an agricultural watershed (Guadiana River, SW Spain). *Ecological Indicators*, **9**, 693–703.

van Dam H., Mertens A. & Sinkeldam J. (1994) A coded checklist and ecological indicator values of freshwater diatoms from the Netherlands. *Netherlands Journal of Aquatic Ecology*, **28**, 117–133.

Venäläinen A. & Heikinheimo M. (2002) Meteorological data for agricultural applications. *Physics and Chemistry of the Earth*, **27**, 1045–1050.

Verleyen E., Hodgson D.A., Vyverman W., Roberts D., McMinn A., Vanhoutte K. *et al.* (2003) Modelling diatom responses to climate induced fluctuations in the moisture balance in continental Antarctic lakes. *Journal of Paleolimnology*, **30**, 195–215.

Verleyen E., Vyverman W., Sterken M., Hodgson D.A., De Wever A., Juggins S. *et al.* (2009) The importance of dispersal related and local factors in shaping the taxonomic structure of diatom metacommunities. *Oikos*, **118**, 1239–1249.

Vyverman W. & Sabbe K. (1995) Diatom-temperature transfer functions based on the altitudinal zonation of diatom assemblages in Papua New Guinea: a possible tool in the reconstruction in regional palaeoclimatic changes. *Journal of Paleolimnology*, **13**, 65–77.

Vyverman W., Verleyen E., Sabbe K., Vanhoutte K., Sterken M., Hodgson D.A. *et al.* (2007) Historical processes constrain patterns in global diatom diversity. *Ecology*, **88**, 1924–1931.

Weckström J., Korhola A. & Blom T. (1997) The relationship between diatoms and water temperature in thirty subarctic Fennoscandian lakes. *Arctic and Alpine Research*, **29**, 75–92.

Winter J. & Duthie H. (2000) Epilithic diatoms as indicators of stream total N and total P concentration. *Journal of the North American Benthological Society*, **19**, 32–49.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. The results of principal component analysis (PCA) showing the interrelations of all the available variables.

Fig. S2. The bivariate matrix for modelled variables.

Fig. S3. The ordination plot of redundancy analysis (RDA) on the species data of all the available variables.

Fig. S4. Results of partial redundancy analysis (pRDA) for species abundance data among local environmental and climatic variable groups.

Fig. S5. Spatial autocorrelation as spatial correlograms for modelled variables.

Fig. S6. Relationships between observed and diatom inferred values.

Fig. S7. Responses of selected species to conductivity and growing degree days.

Table S1. Summary table of all the available variables.

Table S2. Summary of principal component analysis of all the available variables.

Table S3. The results of redundancy analysis (RDA) on the species data of all the available variables.

(Manuscript accepted 11 February 2016)